2009 ISSCC

The New Era of Scaling in an SoC World

Mark Bohr

Intel Senior Fellow

Logic Technology Development

The End of Scaling is Near?

"Optical lithography will reach its limits in the range of 0.75-0.50 microns"

"Minimum geometries will saturate in the range of 0.3 to 0.5 microns"

"X-ray lithography will be needed below 1 micron"

"Minimum gate oxide thickness is limited to ~2 nm"

"Copper interconnects will never work"

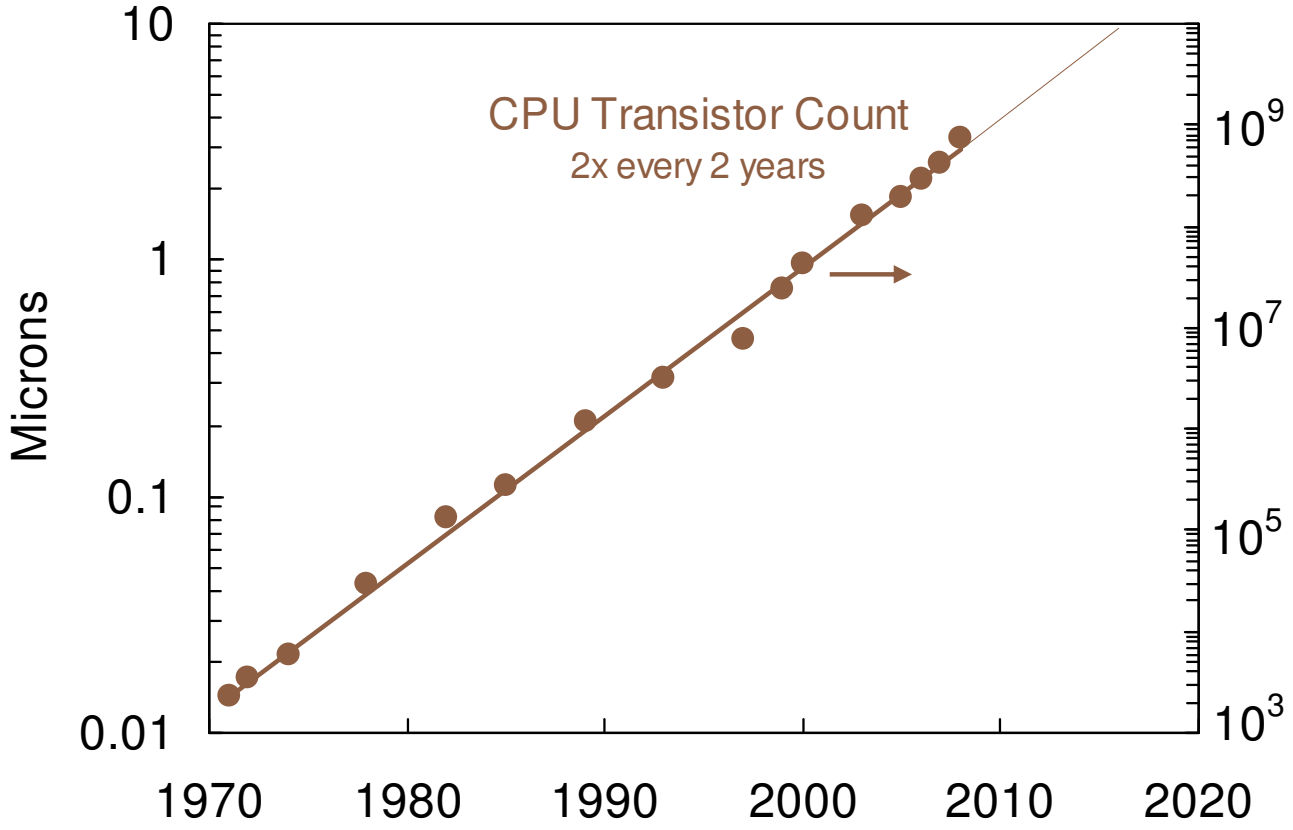
"Scaling will end in ~10 years"

Perceived barriers are meant to be surmounted, circumvented or tunneled through

Outline

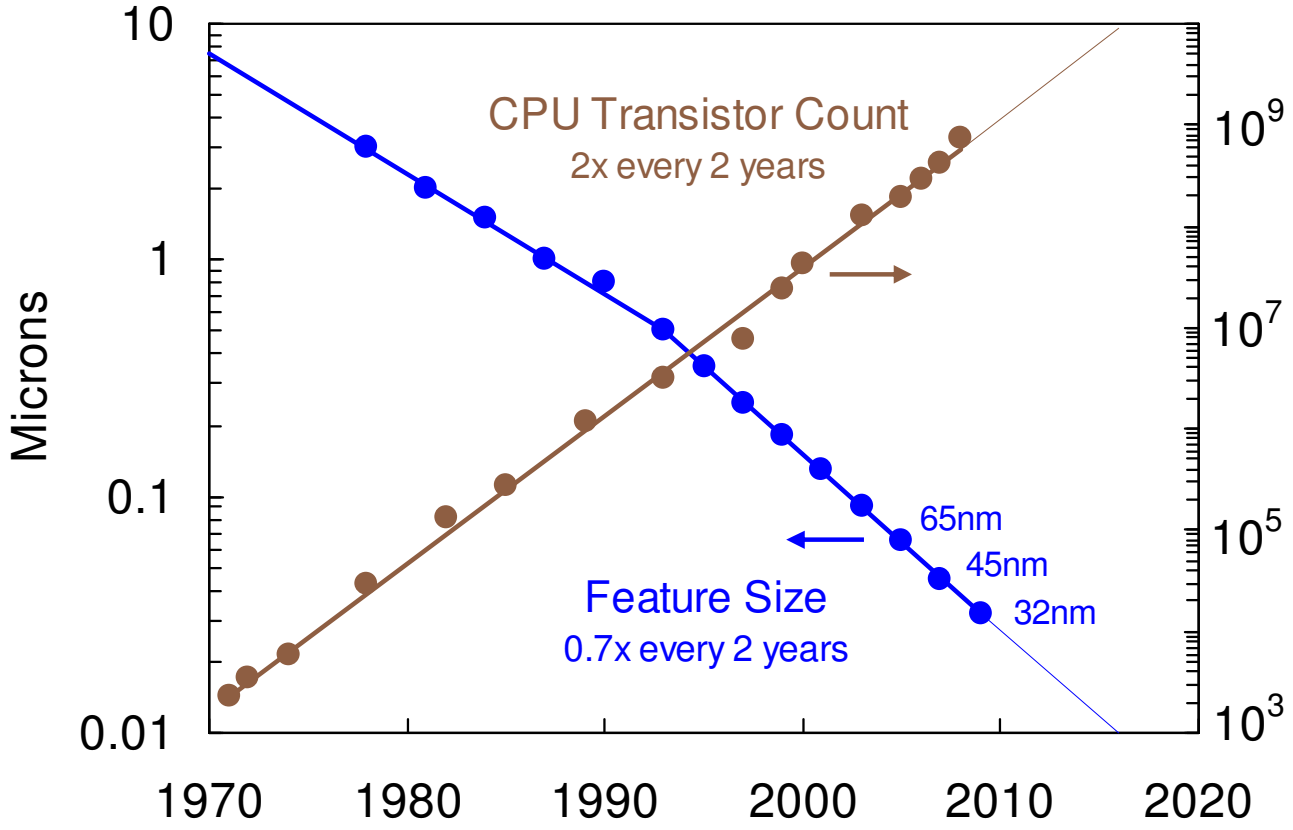
- Transistor Scaling
- Microprocessor Evolution
- Vision of the Future

Scaling Trends



Transistor dimensions scale to improve performance, reduce power and reduce cost per transistor

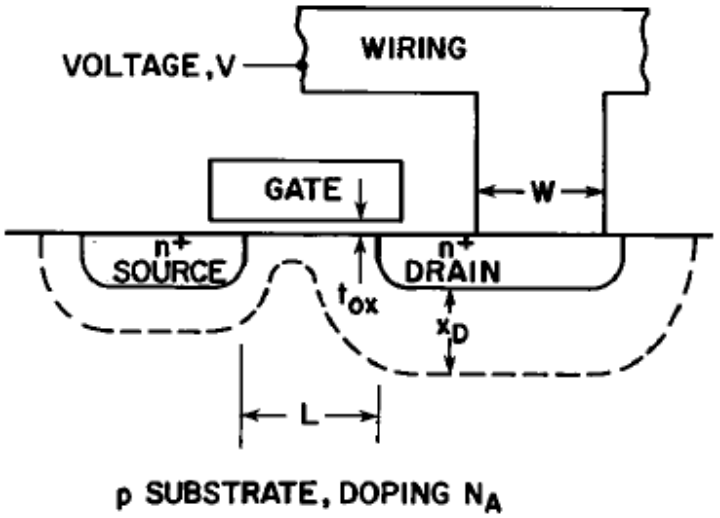
Scaling Trends



Transistor dimensions scale to improve performance, reduce power and reduce cost per transistor

MOSFET Scaling

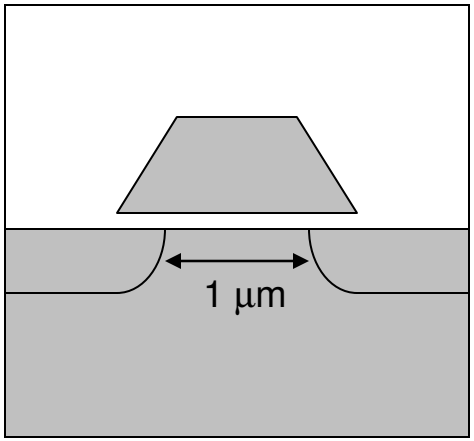
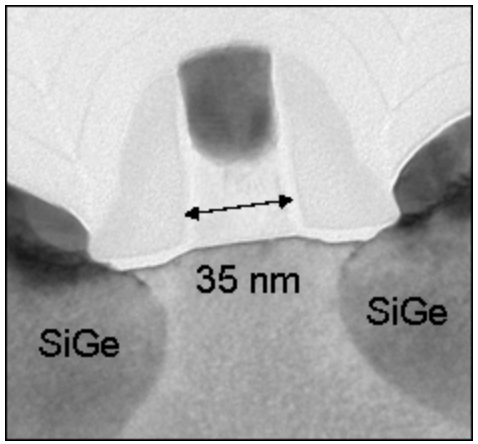
Device or Circuit Parameter	Scaling Factor
Device dimension *tox, L, W*	1/κ
Doping concentration *Na*	κ
Voltage *V*	1/κ
Current *I*	1/κ
Capacitance *εA/t*	1/κ
Delay time/circuit *VC/I*	1/κ
Power dissipation/circuit *VI*	$1/\kappa^2$
Power density *VI/A*	1



R. Dennard, IEEE JSSC, 1974

Classical MOSFET scaling was first described in 1974

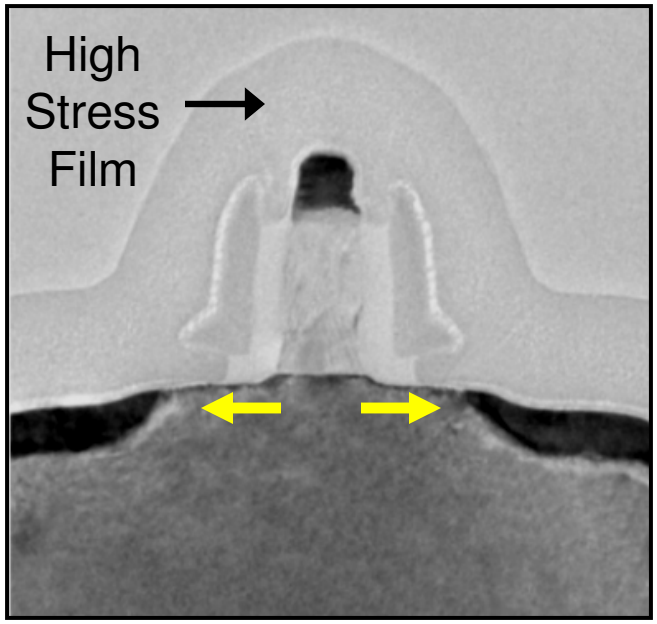
30 Years of MOSFET Scaling

	Dennard 1974	Intel 2005
		
Gate Length:	1.0 μm	35 nm
Gate Oxide Thickness:	35 nm	1.2 nm
Operating Voltage:	4.0 V	1.2 V

Classical scaling ended in the early 2000s due to gate oxide leakage limits

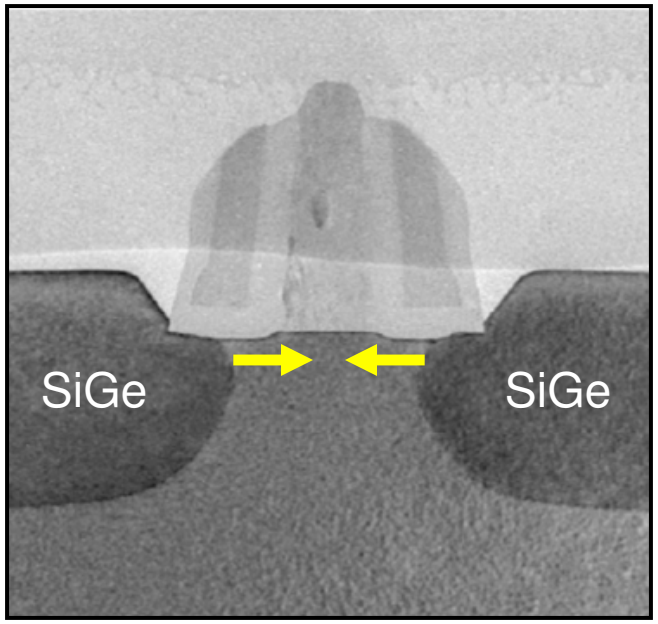
90 nm Strained Silicon Transistors

NMOS



SiN cap layer
Tensile channel strain

PMOS

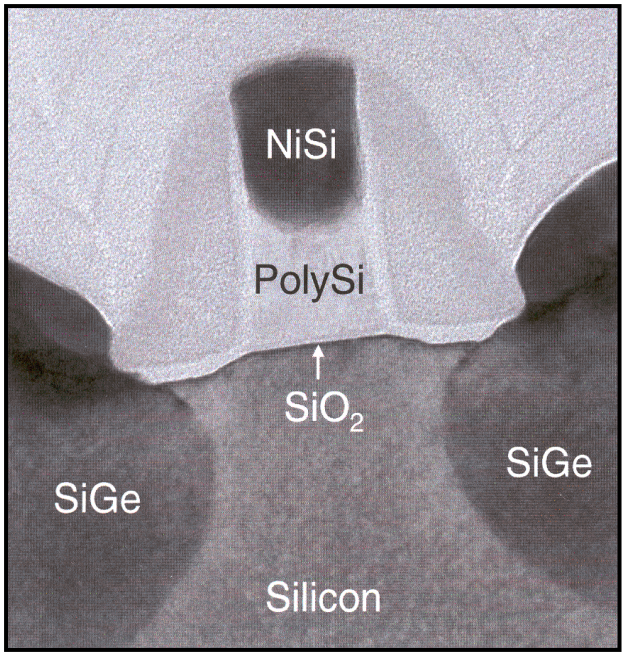


SiGe source-drain
Compressive channel strain

Strained silicon provided increased drive currents, making up for lack of gate oxide scaling

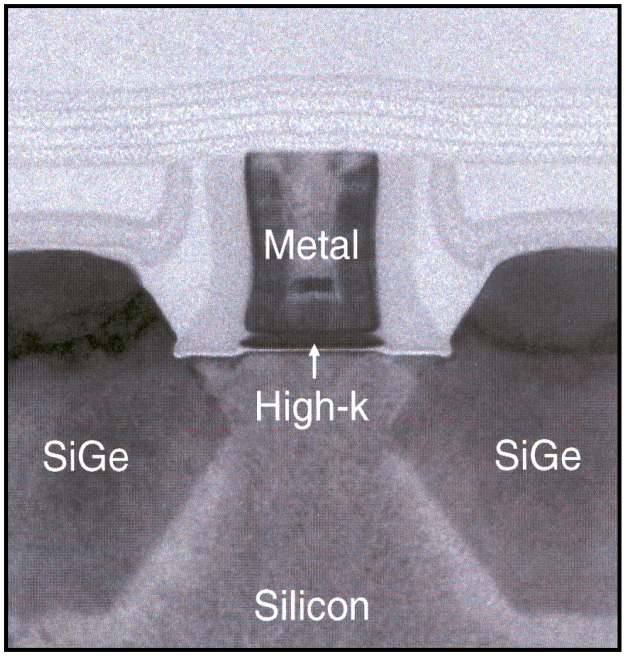
High-k + Metal Gate Transistors

65 nm Transistor



SiO_2 dielectric
Polysilicon gate electrode

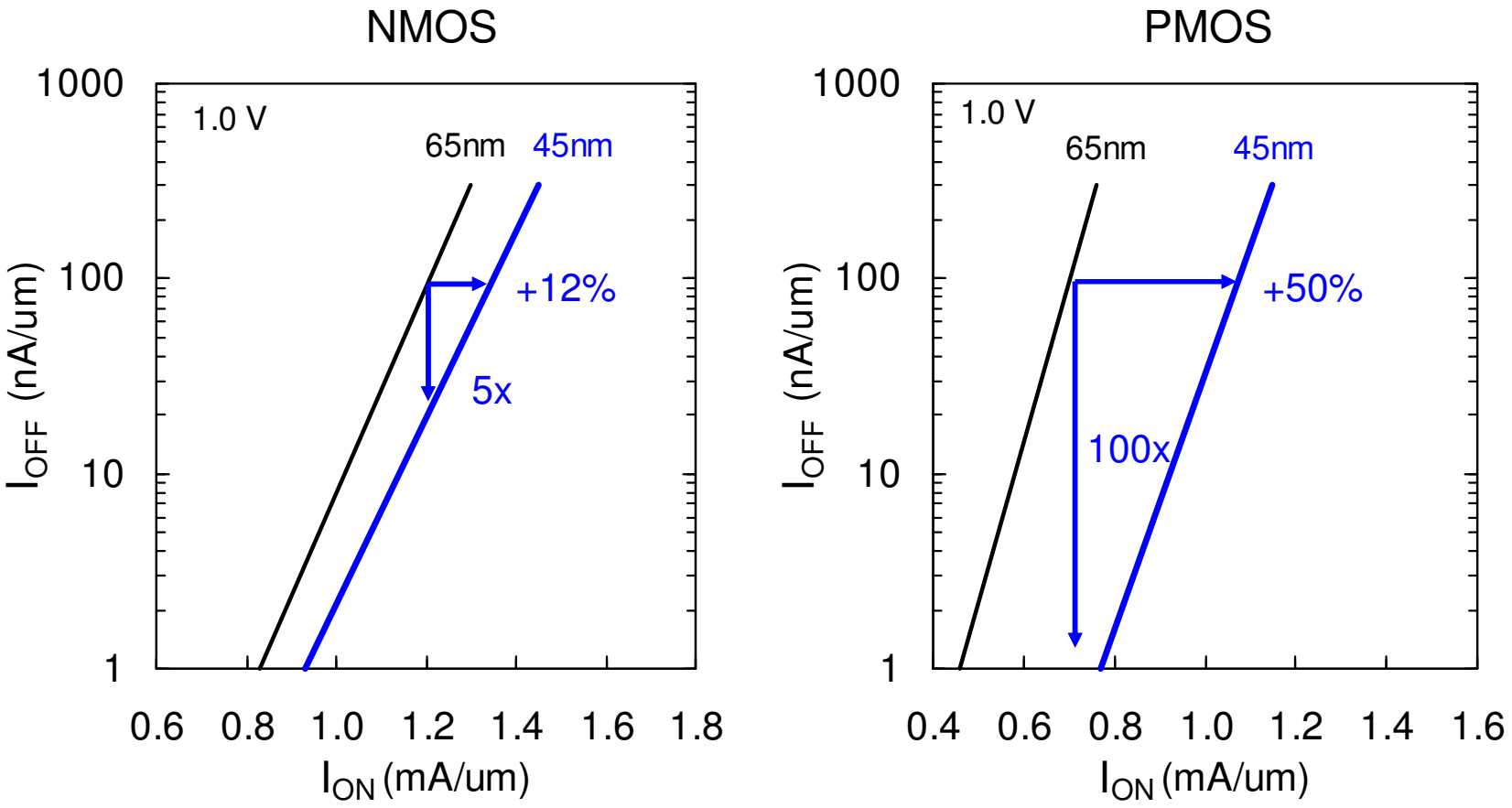
45 nm HK+MG



Hafnium-based dielectric
Metal gate electrode

High-k + metal gate transistors
break through gate oxide scaling barrier

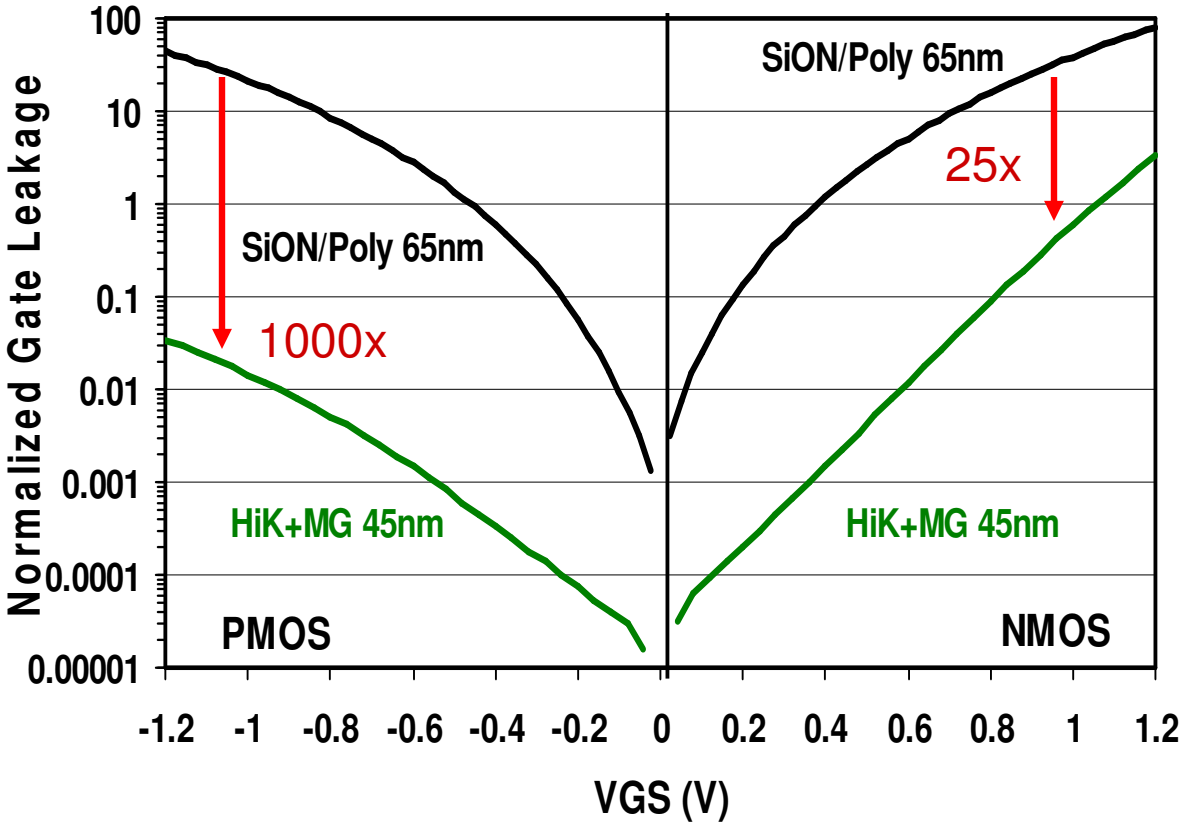
Transistor Performance Increase



45 nm HK+MG provides average 30% drive current increase or >5x I_{OFF} leakage reduction

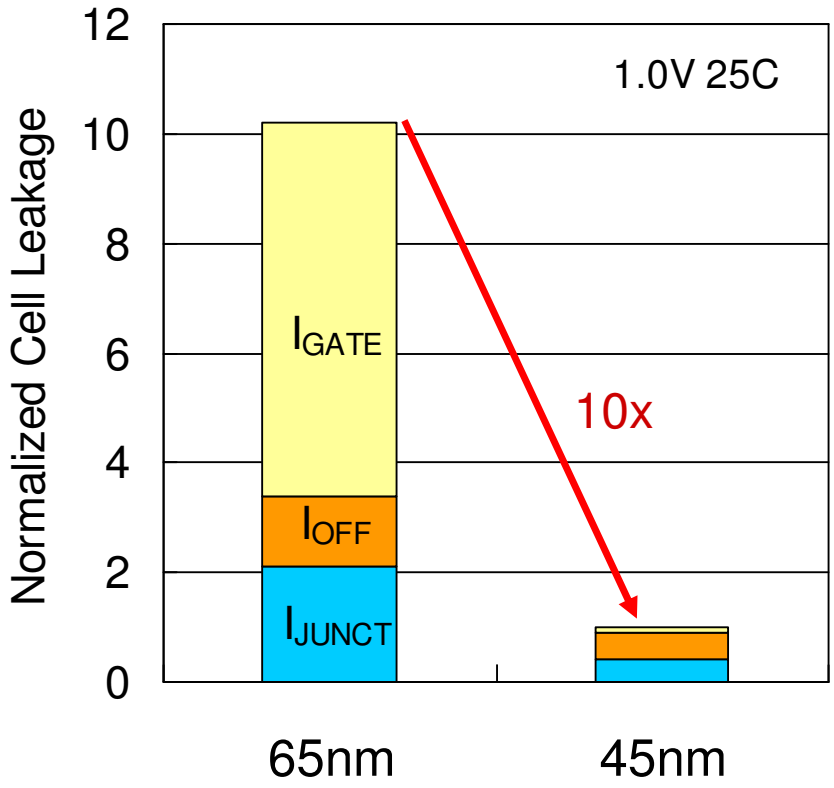
Ref. K. Mistry, IEDM '07

Gate Leakage Reduction



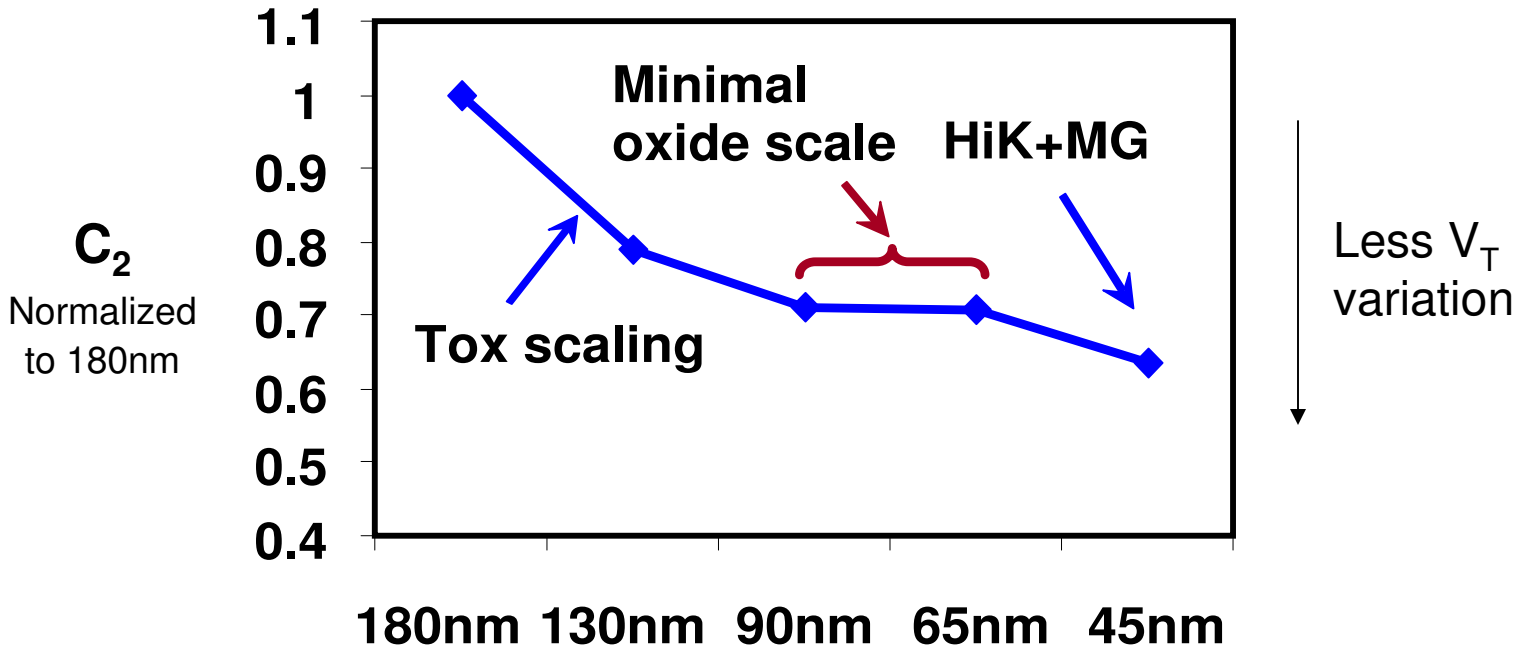
HK+MG significantly reduces gate leakage

Bitcell Leakage Reduction



SRAM bitcell leakage reduced ~10x

V_T Variability Reduction

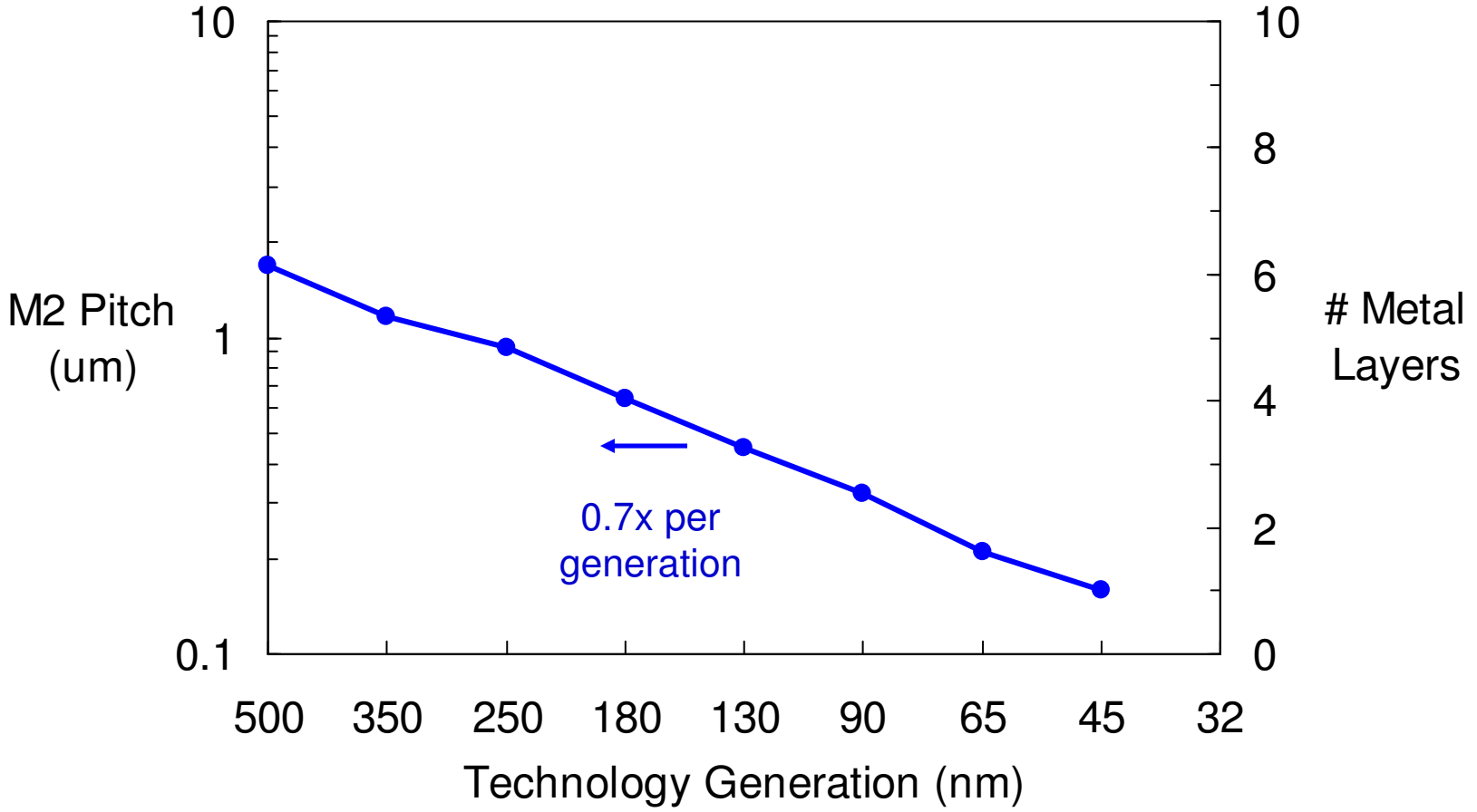


$$\sigma V_{Tran} = \left(\frac{\sqrt[4]{4q^3 \varepsilon_{si} \phi_B}}{2} \right) \cdot \frac{T_{ox}}{\varepsilon_{ox}} \cdot \left(\frac{\sqrt[4]{N}}{\sqrt{Leff \cdot Zeff}} \right) = \frac{1}{\sqrt{2}} \left(\frac{C_2}{\sqrt{Leff \cdot Zeff}} \right)$$

HK+MG provides oxide scaling needed for variability reduction

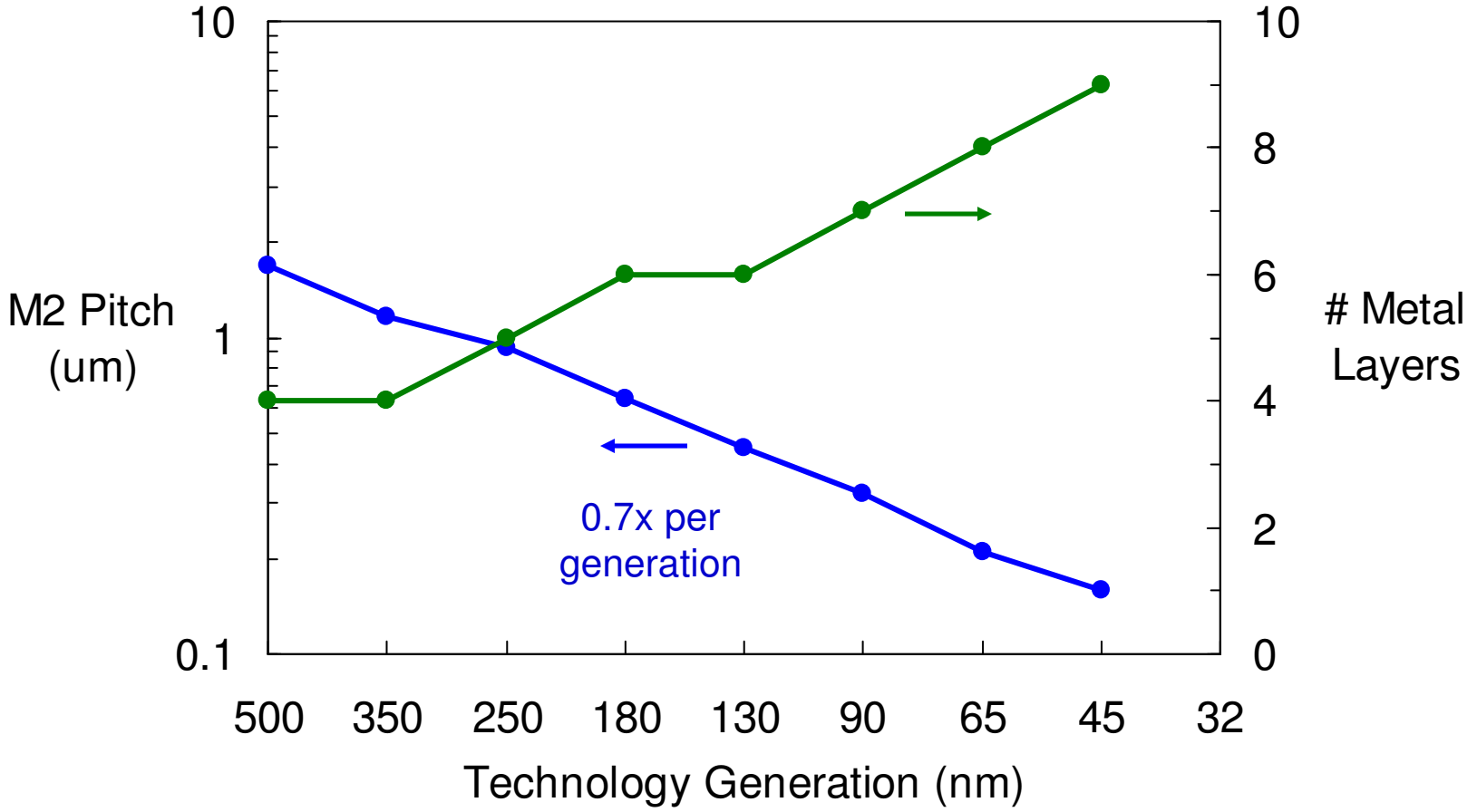
Ref. K. Kuhn, IEDM '07

Interconnect Trends



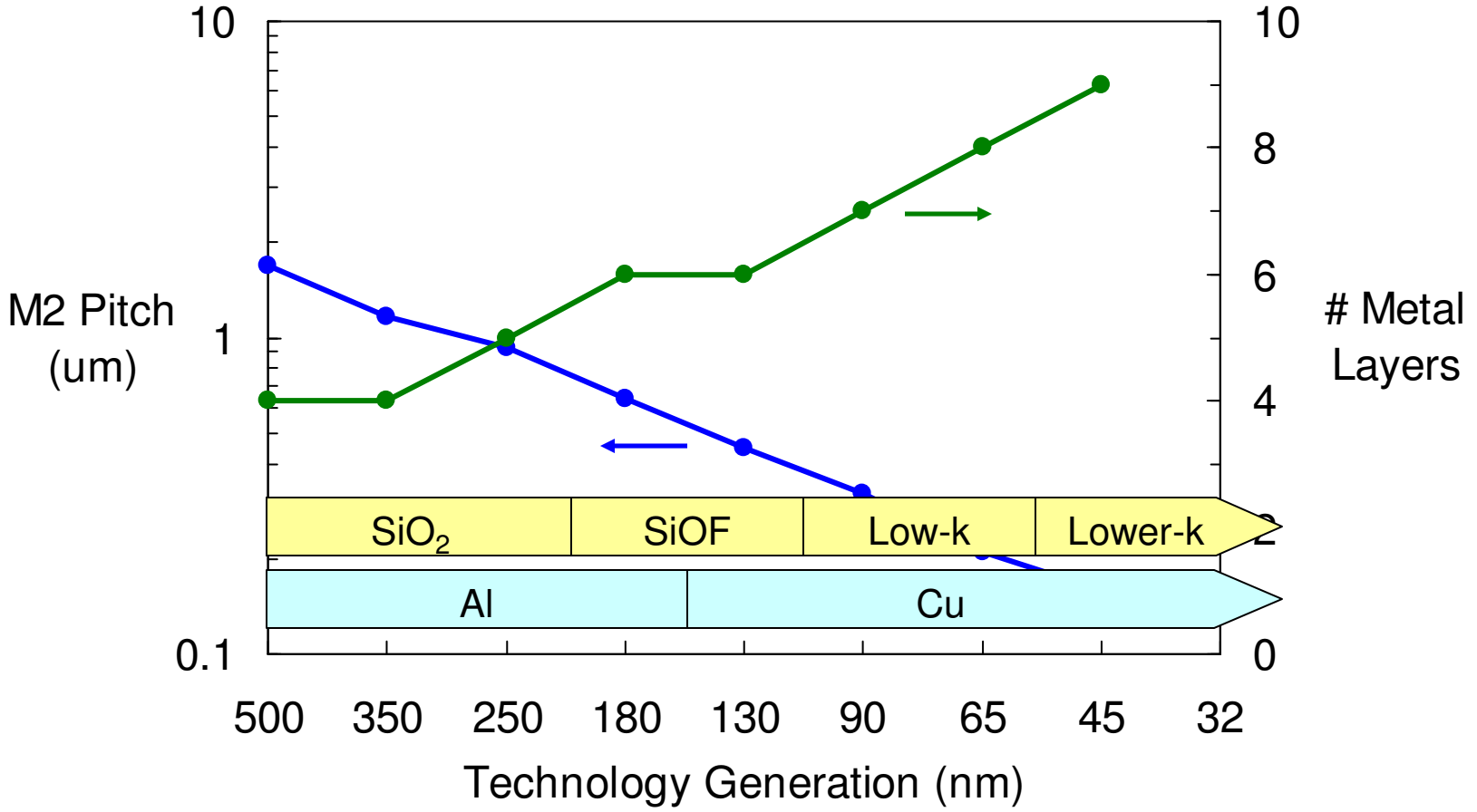
Added metal layers + material improvements enable interconnect scaling

Interconnect Trends



Added metal layers + material improvements enable interconnect scaling

Interconnect Trends



Added metal layers + material improvements enable interconnect scaling

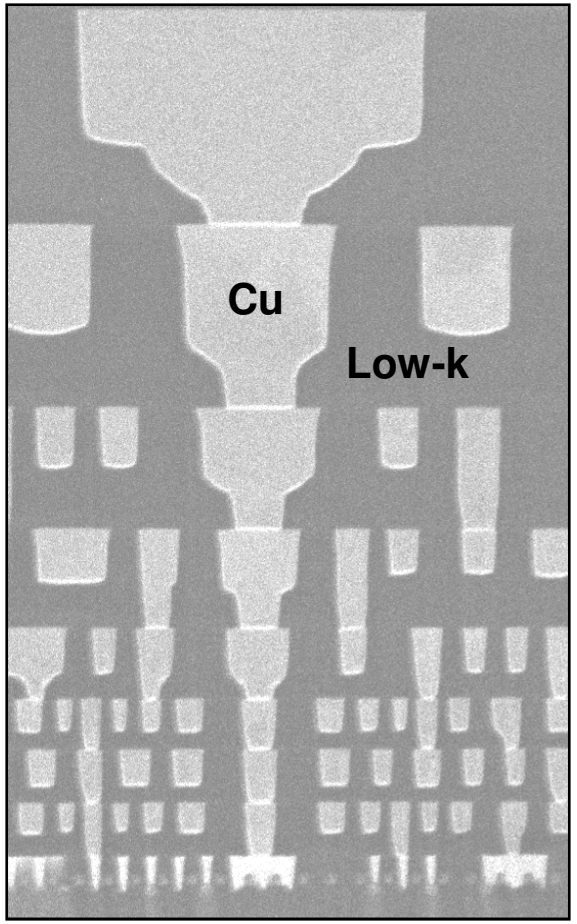
45 nm Interconnects

Loose pitch + thick metal on upper layers

- High speed global wires
- Low resistance power grid

Tight pitch on lower layers

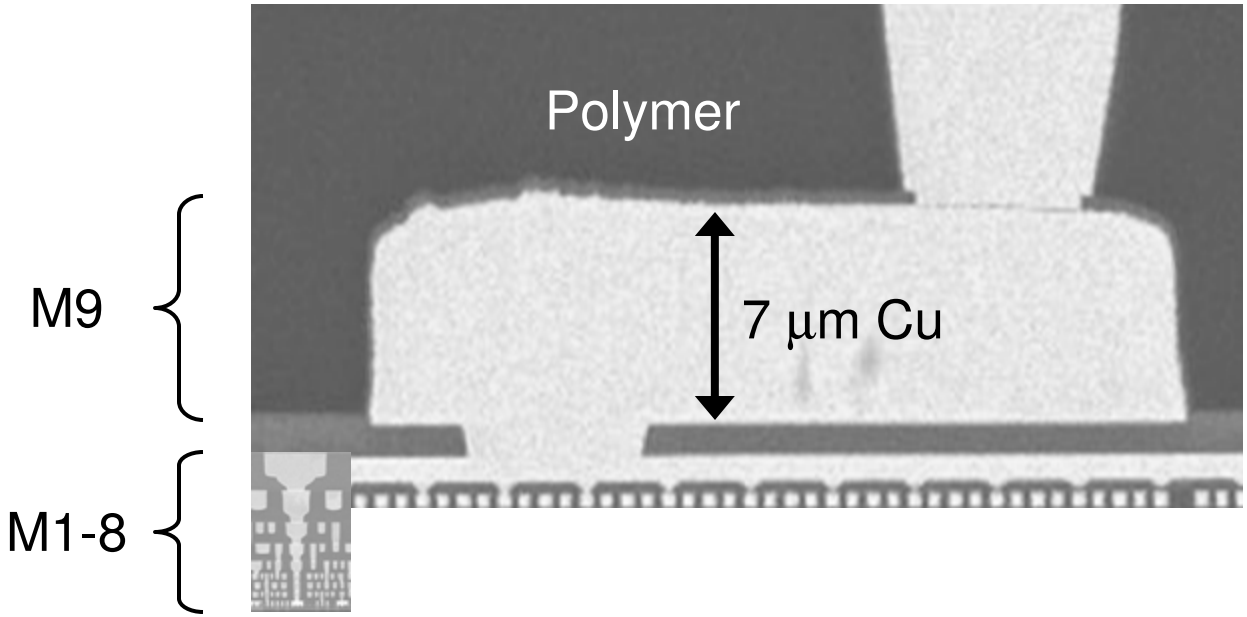
- Maximum density for local interconnects



	Pitch (nm)
M8	810
M7	560
M6	360
M5	280
M4	240
M3	160
M2	160
M1	160

Hierarchical interconnect pitches

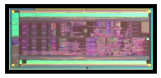
45 nm Interconnects



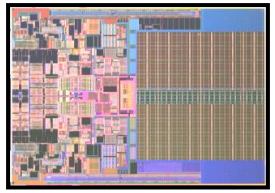
Thick M9 for very low resistance on-die power routing

45 nm Microprocessor Products

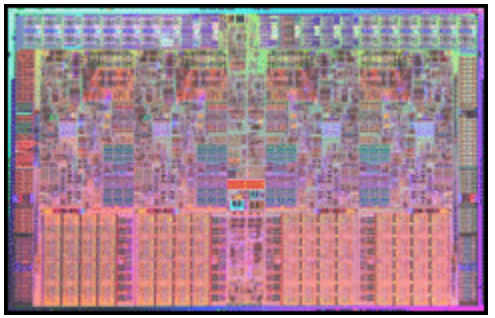
Single Core

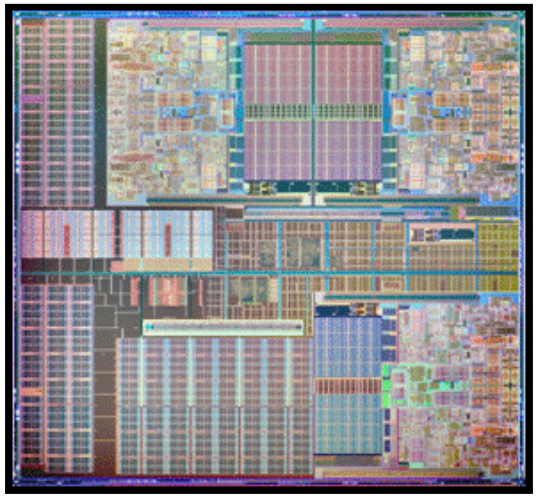


Dual Core

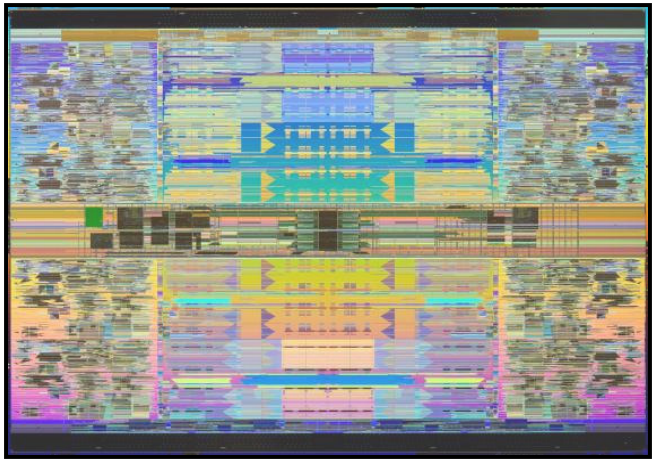


Quad Core





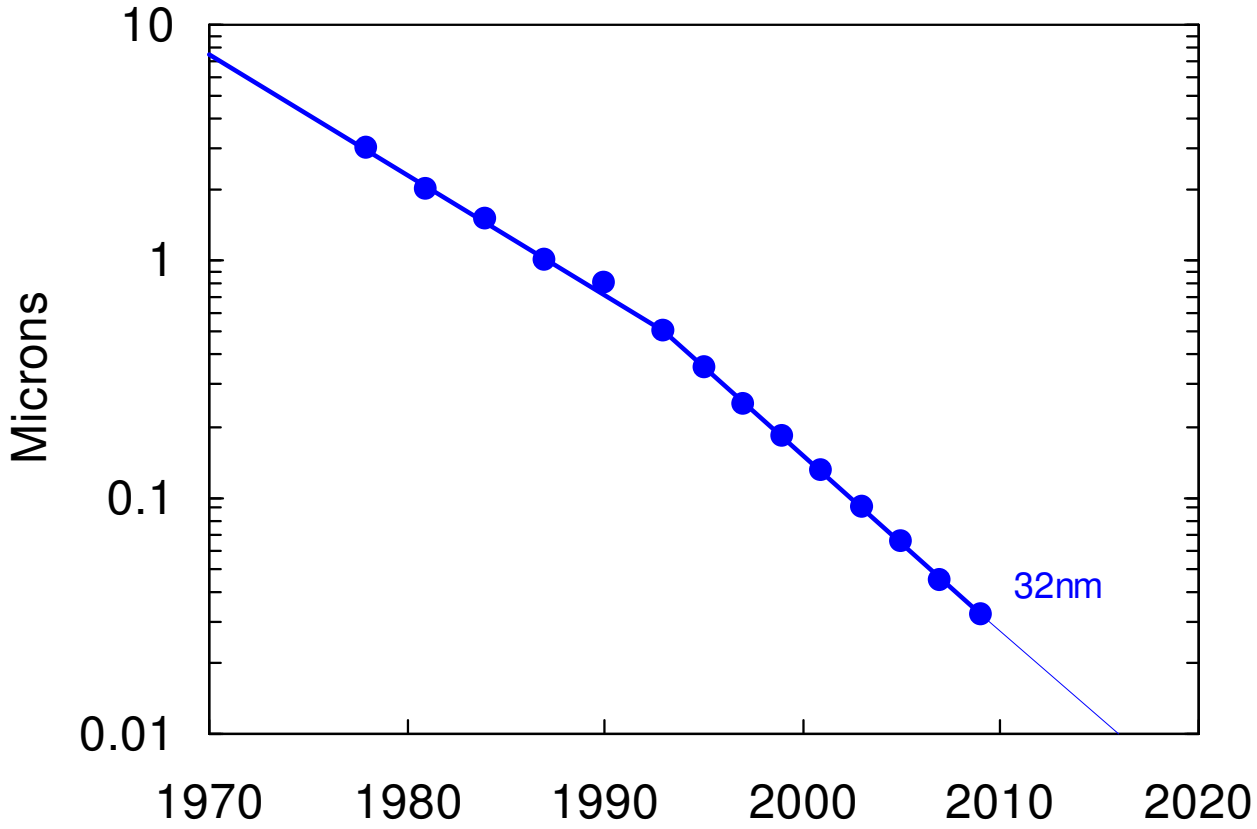
6 Core



8 Core

45 nm process serves microprocessor applications from low power to high performance

32 nm Generation

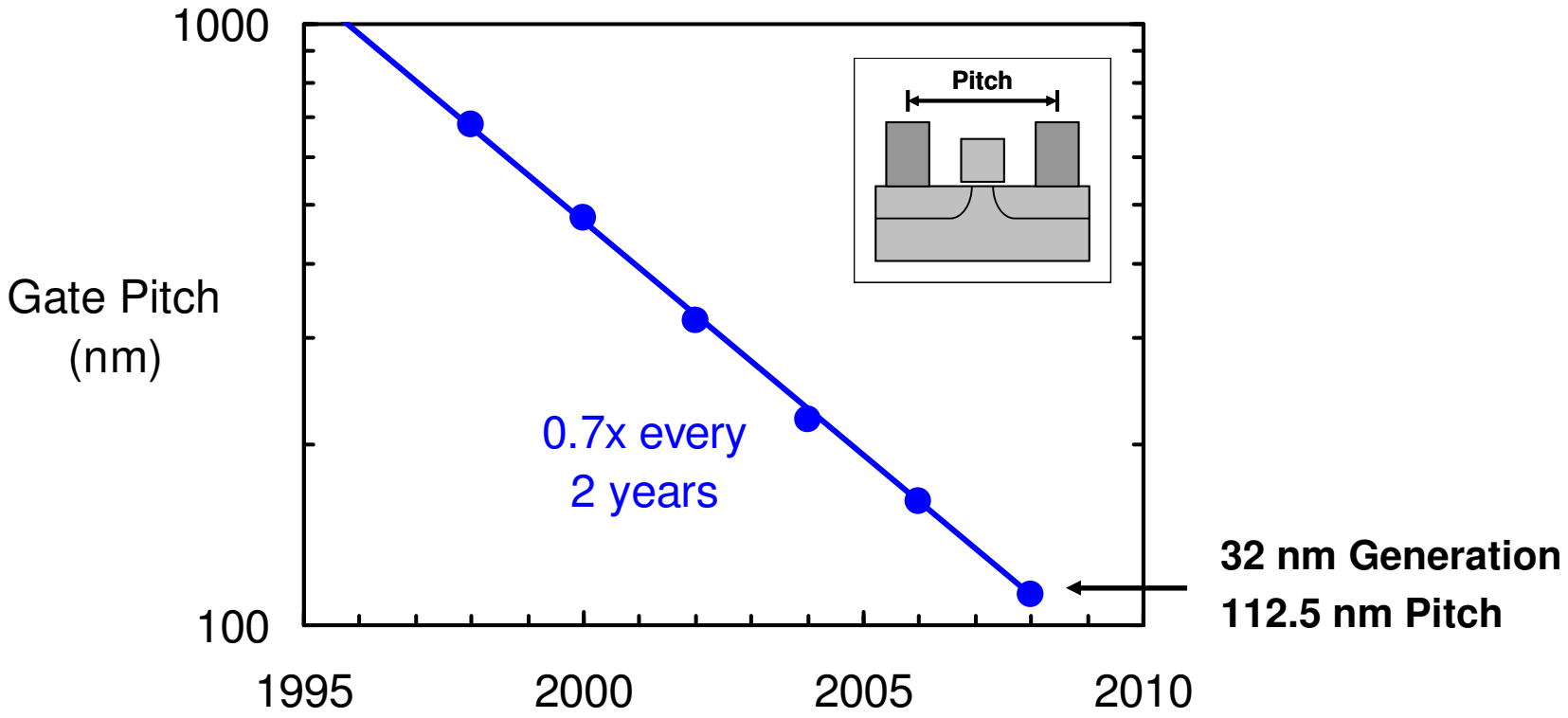


32 nm Logic Technology

- 2nd generation high-k + metal gate transistors
  - High-k EOT scaled from 1.0 nm to 0.9 nm
  - Replacement metal gate process flow
  - 4th generation strained silicon
- 9 copper + low-k interconnect layers
  - Hierarchical interconnect pitches
  - Thick M9 for power routing
- Immersion lithography on critical layers
  - 70% transistor and interconnect pitch scaling
  - 50% SRAM cell area scaling
- Pb-free and halogen-free packages

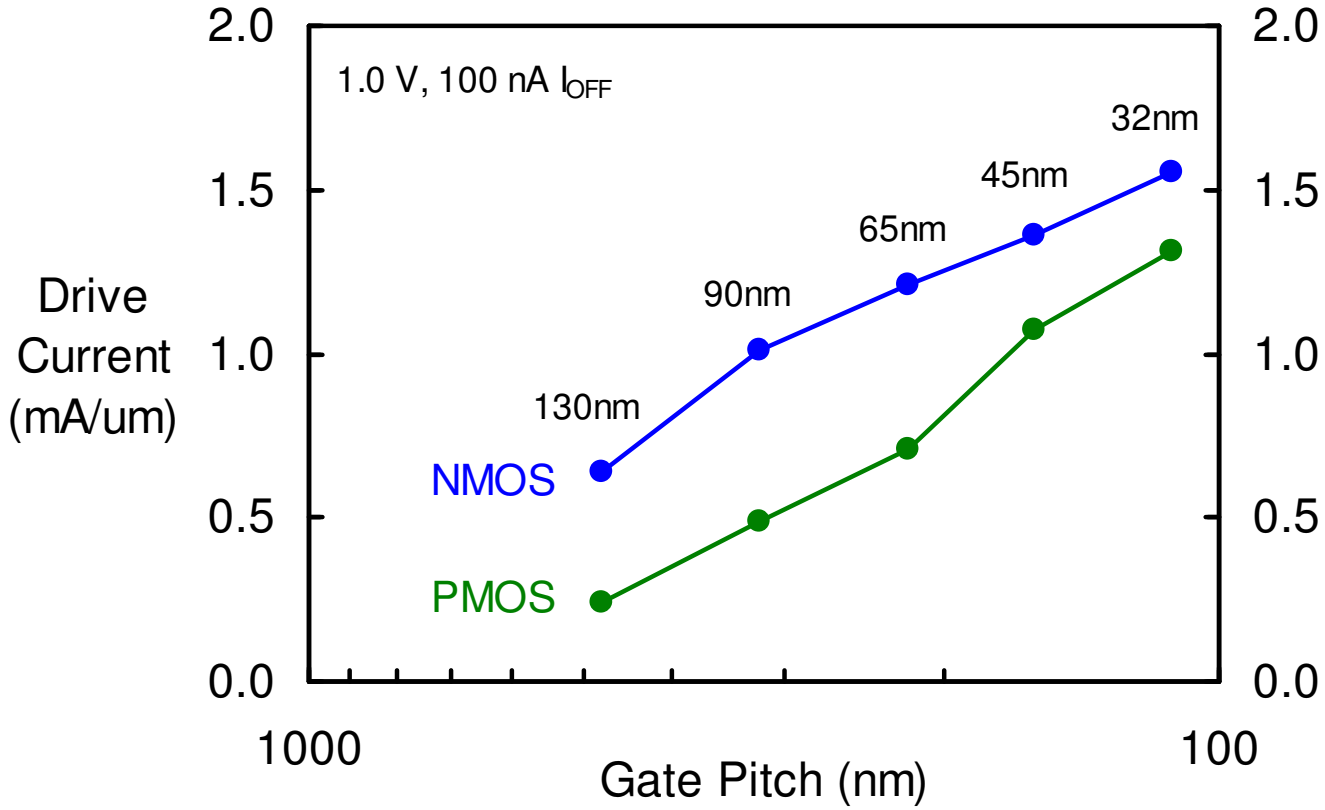
Higher performance, lower power, lower cost per transistor

Contacted Gate Pitch Trend



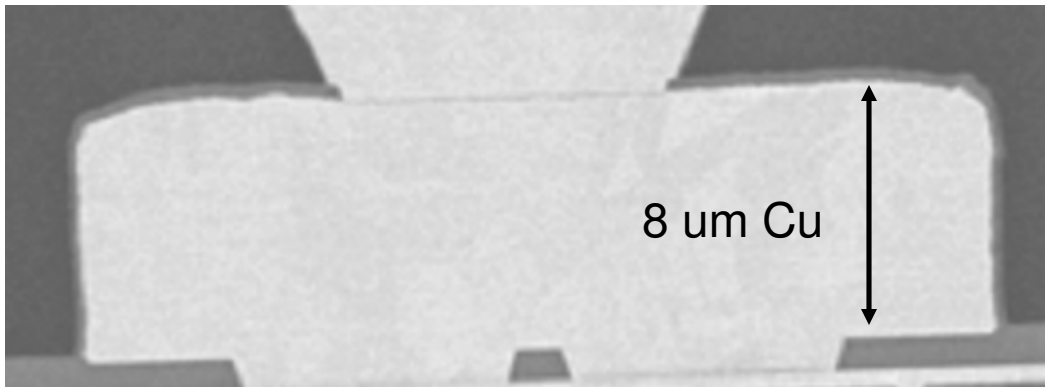
Transistor gate pitch continues to scale 0.7x every 2 years

Transistor Performance

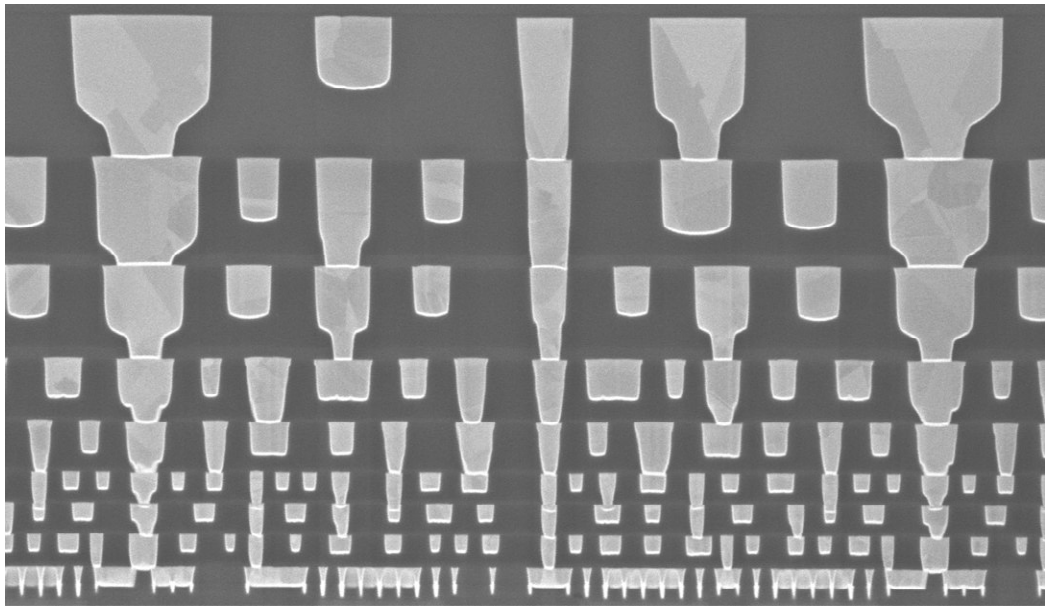


Drive currents continue to increase while gate pitch scales

32 nm Interconnects



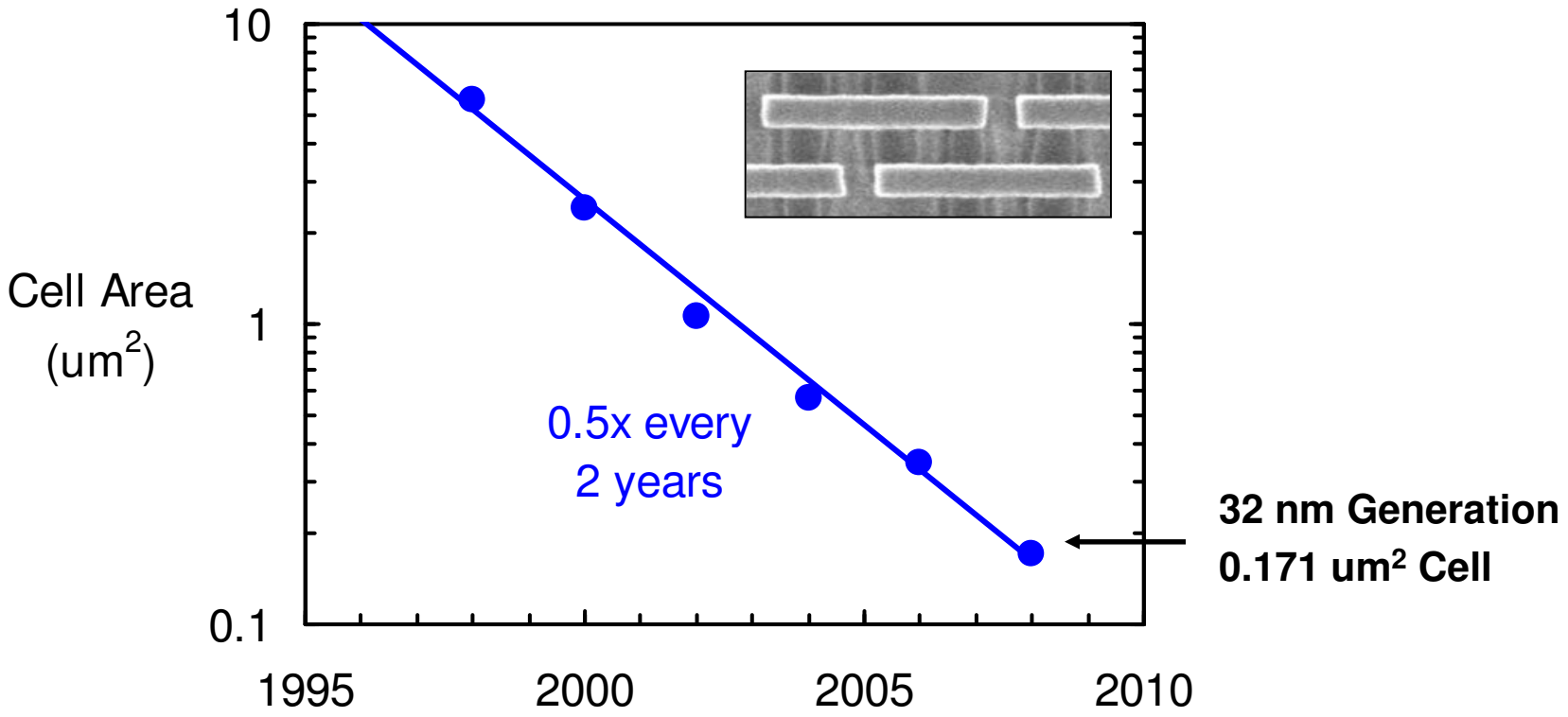
M9



	Pitch (nm)
M8	566.5
M7	450.1
M6	337.6
M5	225.0
M4	168.8
M3	112.5
M2	112.5
M1	112.5

Hierarchical interconnect pitches

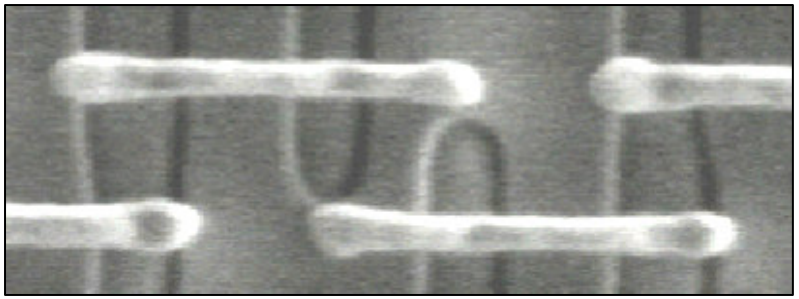
SRAM Cell Size Scaling



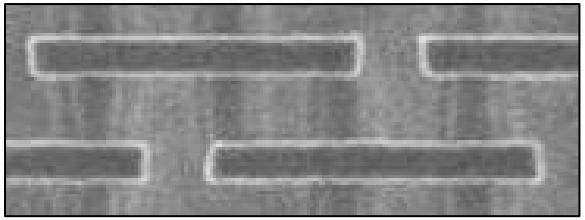
Transistor density continues to double every 2 years

SRAM Cell Scaling

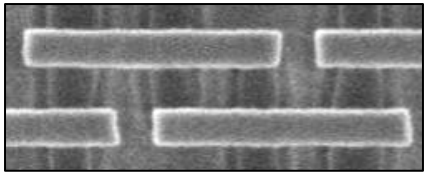
65 nm
0.570 μm^2



45 nm
0.346 μm^2



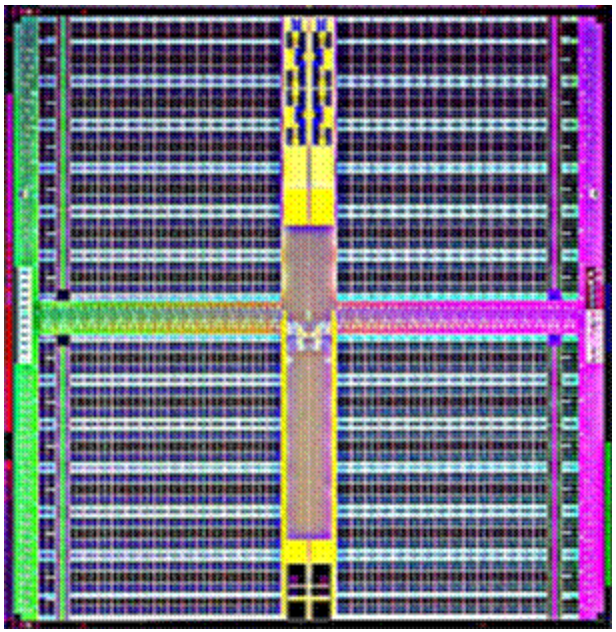
32 nm
0.171 μm^2



Good pattern resolution while scaling feature size and continuing with 193 nm exposure wavelength

32 nm SRAM Test Chip

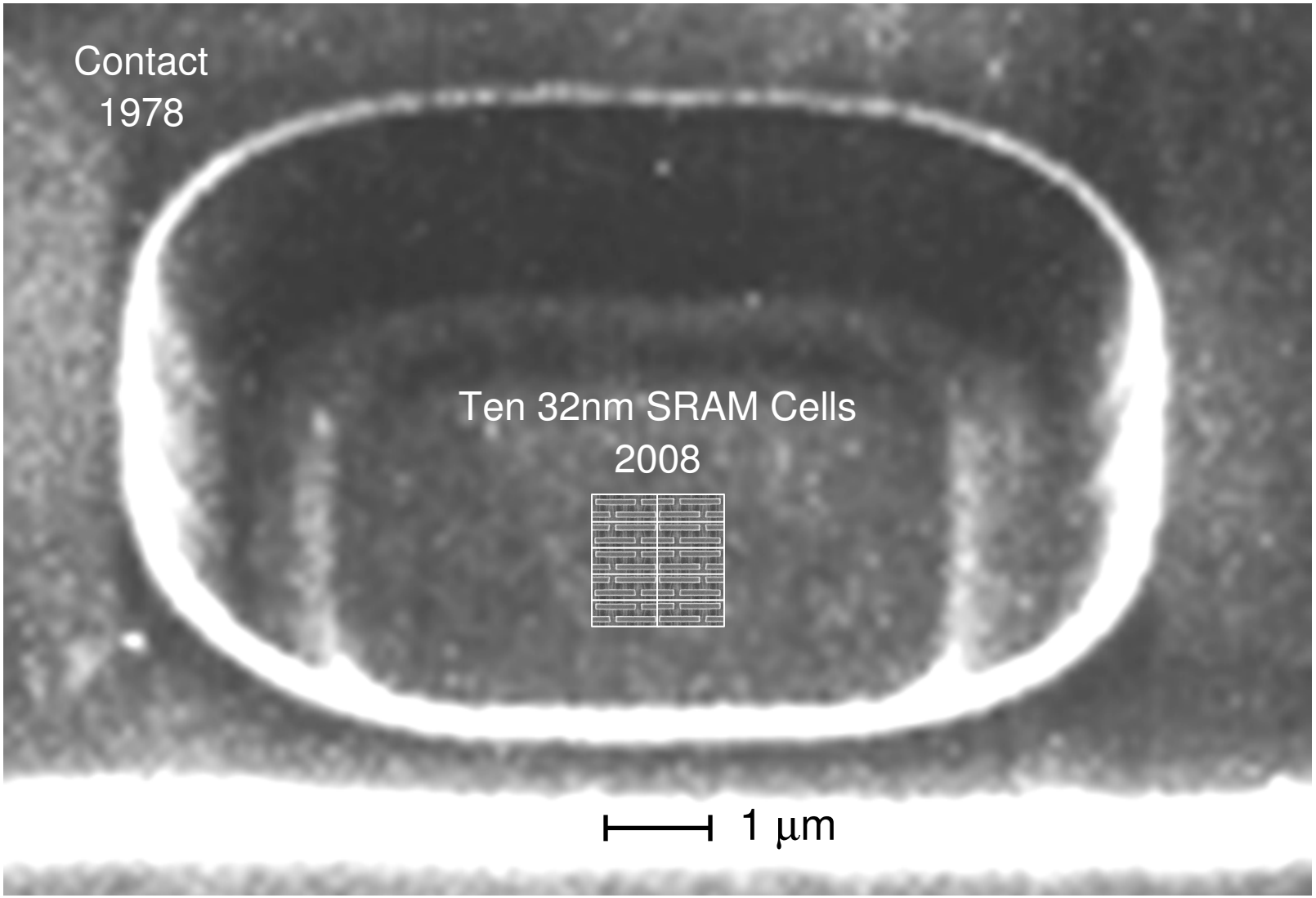
- 291 Mbit
- 0.171 um^2 cell size
- >1.9 billion transistors
- >3.8 GHz operation
- Functional silicon in Aug '07



32 nm SRAM test vehicle included all transistor and interconnect features used on 32 nm microprocessors

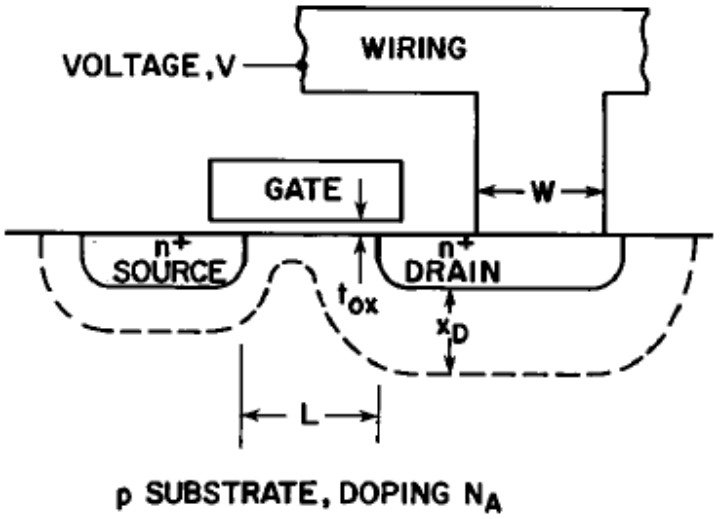
Ref. Y. Wang, paper 27.1, ISSCC '09

30 Years of Scaling



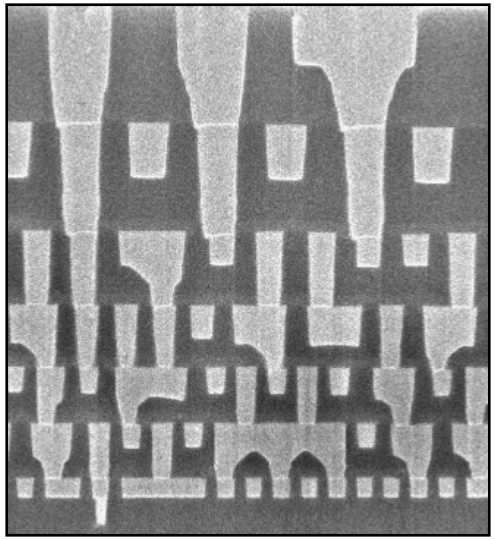
The Old Era of Device Scaling

Device or Circuit Parameter	Scaling Factor
Device dimension *tox, L, W*	1/κ
Doping concentration *Na*	κ
Voltage *V*	1/κ
Current *I*	1/κ
Capacitance *εA/t*	1/κ
Delay time/circuit *VC/I*	1/κ
Power dissipation/circuit *VI*	$1/\kappa^2$
Power density *VI/A*	1

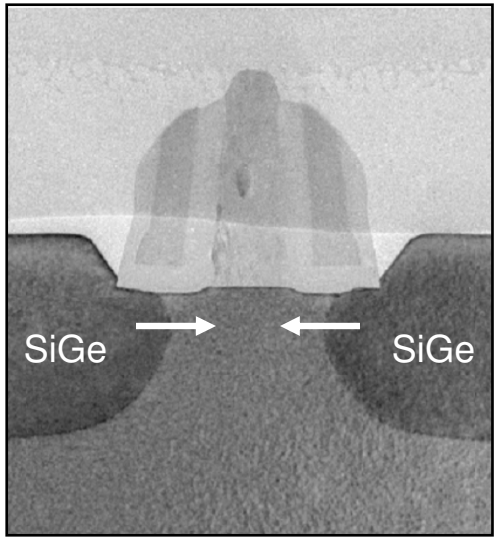


It has served us well for >30 years

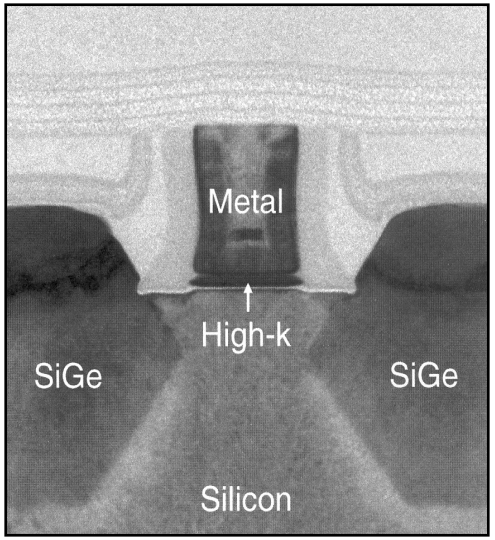
The New Era of Device Scaling



Copper + Low-k



Strained Silicon



High-k + Metal Gate

Modern CMOS scaling is as much about
material and structure innovation as dimensional scaling

Outline

- Transistor Scaling
- Microprocessor Evolution
- Vision of the Future

Microprocessor Evolution

More transistors

Higher frequency

More data bits per cycle

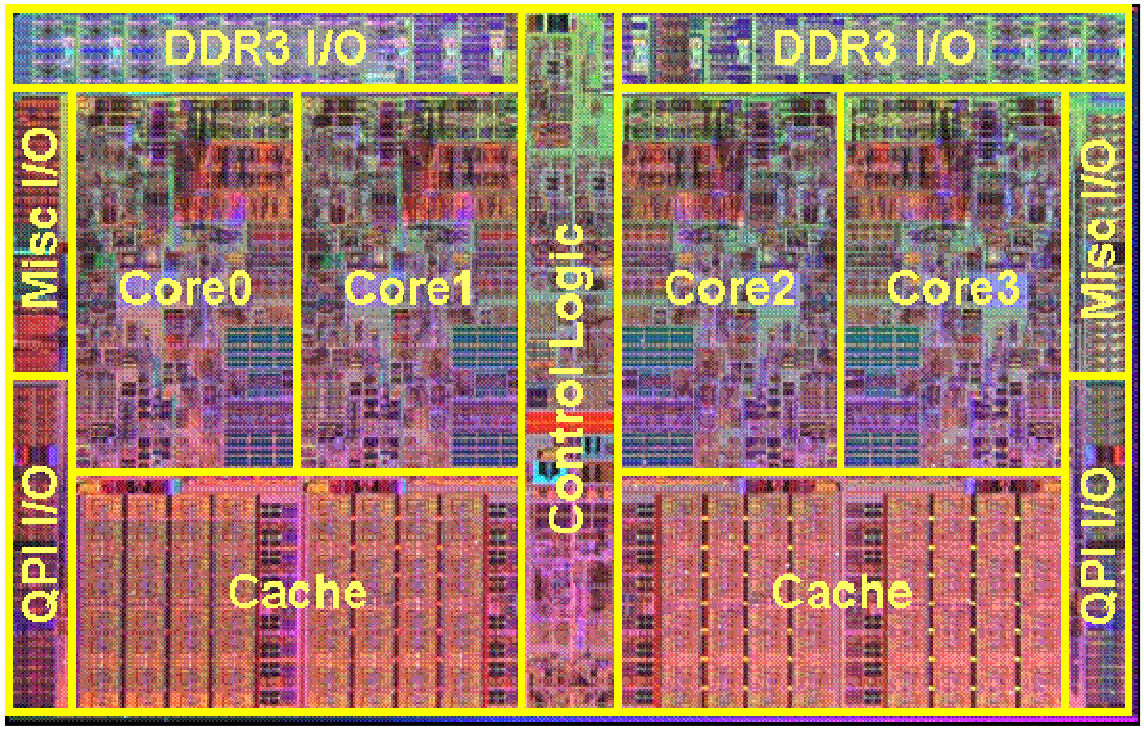
Instruction parallelism

Out-of-order issue

Multi-threading

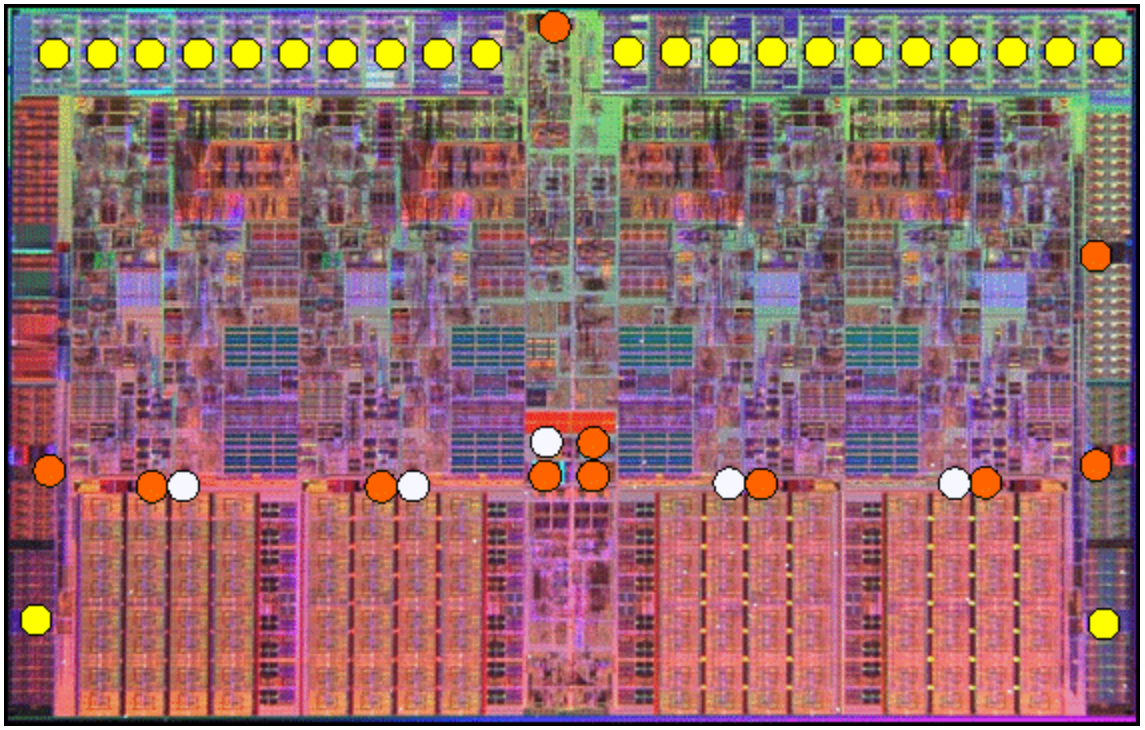
Many of these innovations have been for improved performance, now the challenge is to innovate for power efficiency

45 nm Nehalem CPU



Modern microprocessors are a complex system on a chip with multiple functional units and multiple interfaces

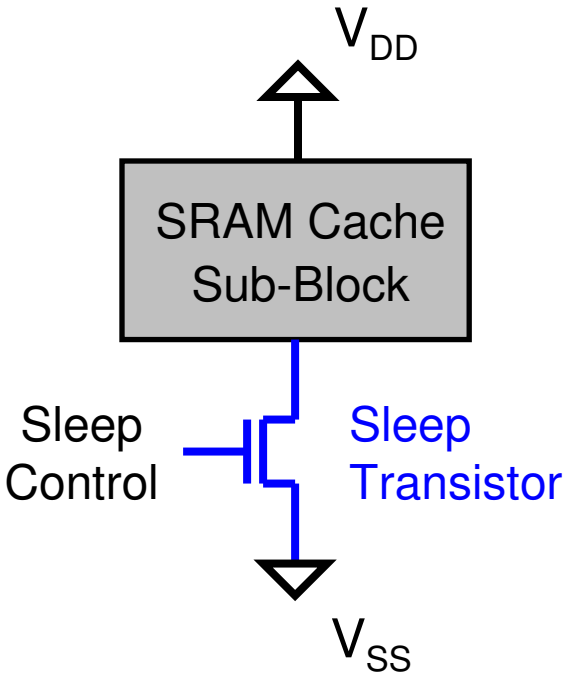
45 nm Nehalem CPU



● 23 master DLL circuits ● 11 PLL circuits ○ 5 digital thermal sensors

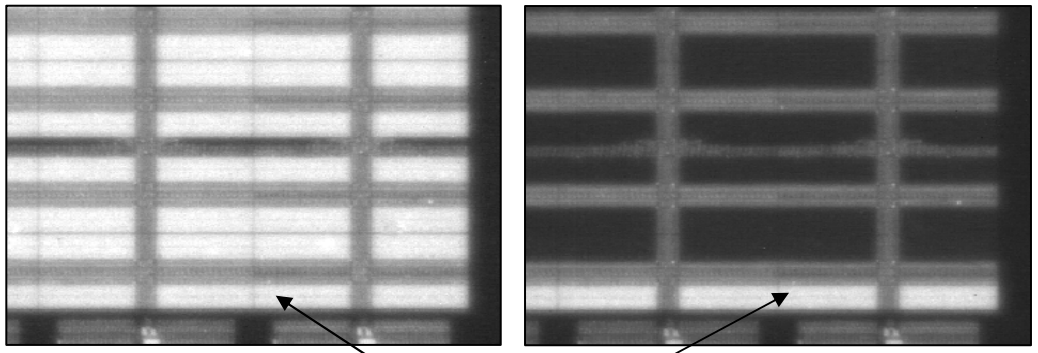
Multiple clocking domains, local control

SRAM Dynamic Sleep Transistors



Normal SRAM
sub-block leakage

Sleep transistors
shut off leakage in
inactive sub-blocks

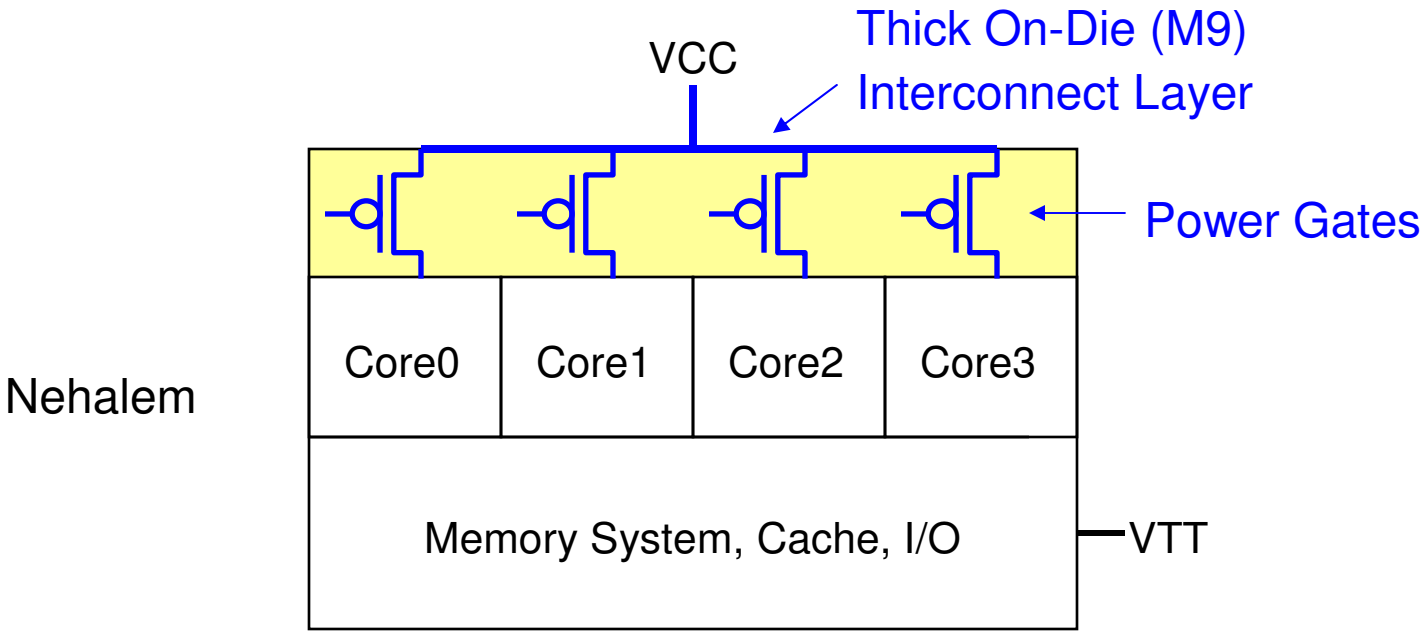


IREM images showing banks being accessed

5-10x leakage reduction during "retention/standby"

Ref. K. Zhang, VLSI Circuits '04

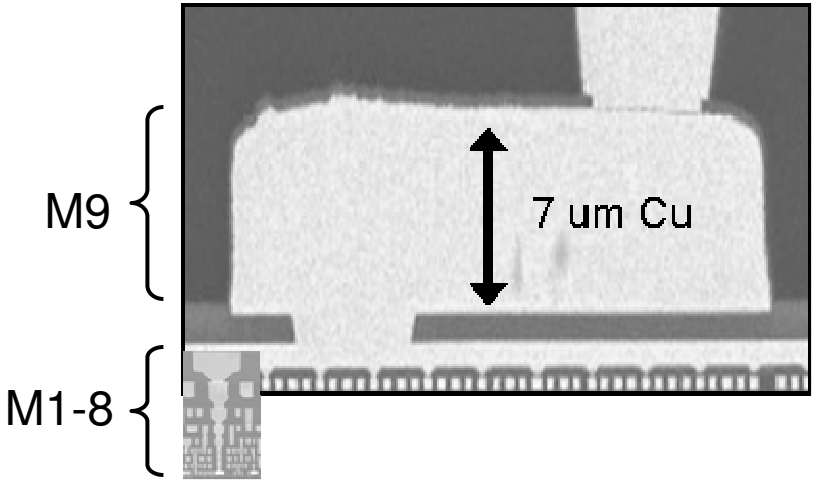
Integrated Power Gates



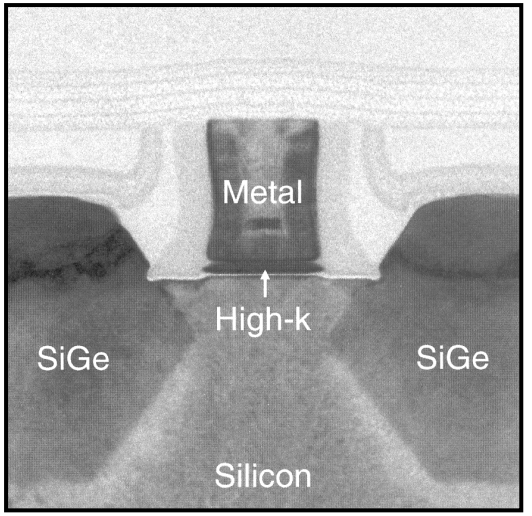
- Shuts off both switching power and leakage power
- Enables idle cores to go to ~0 power, independent of state of other cores on die

Ref. R. Kumar, paper 3.2, ISSCC '09

Power Gates Enabled with Design+Process Co-optimization



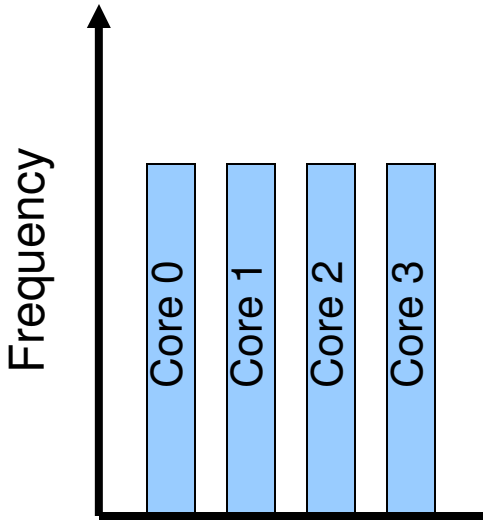
Thick metal 9 layer for low resistance on-die power routing



Ultra-low leakage transistor for high off-resistance power gates

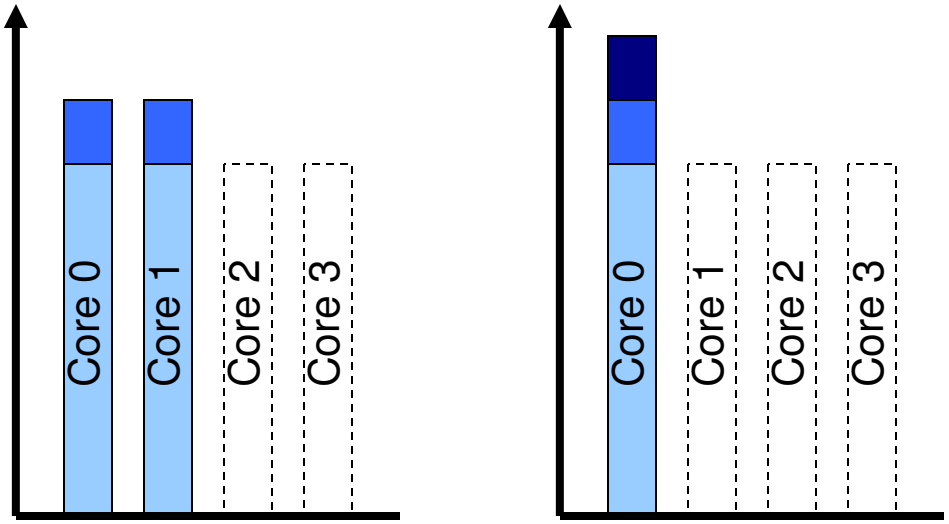
Nehalem Turbo Mode

Many threaded workloads



- All cores operating

Lightly threaded workloads - Turbo Mode

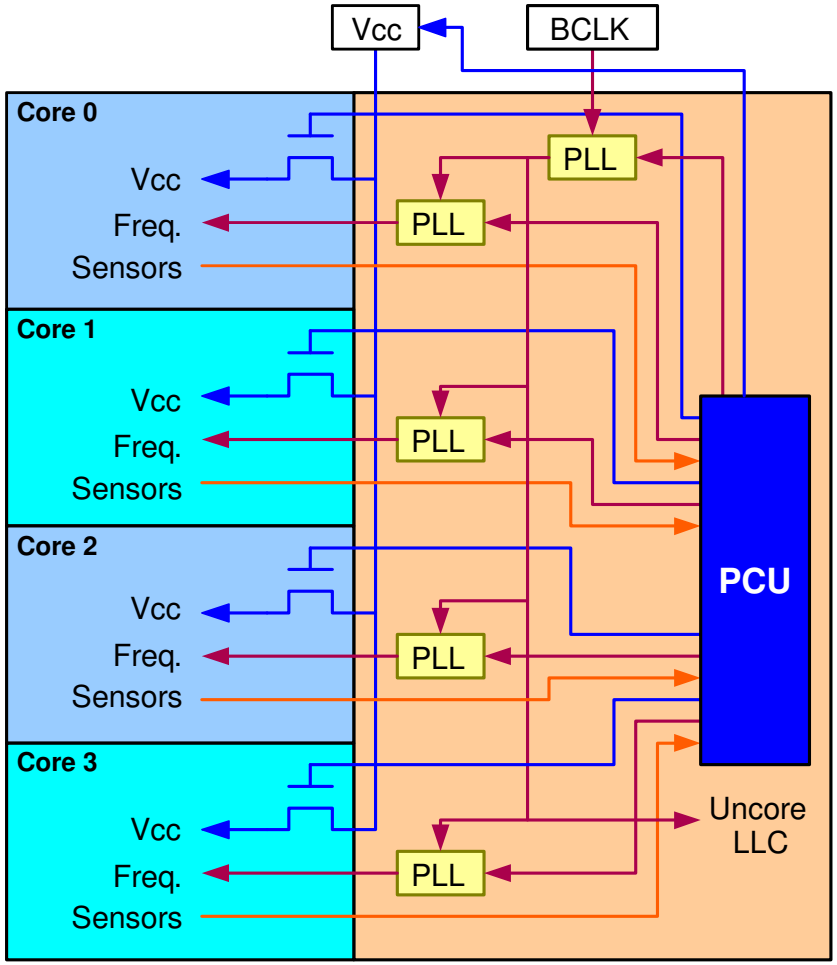


- Power gates shut off some cores
- Zero power for inactive cores
- Higher frequency for active cores

Dynamically delivering optimal performance and energy efficiency

Ref. R. Kumar, paper 3.2, ISSCC '09

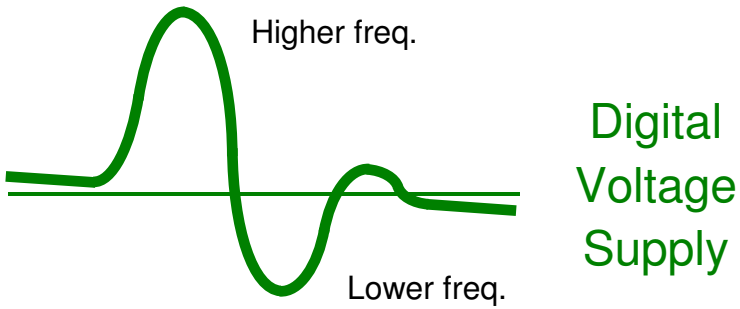
Nehalem Power Control Unit



- Integrated proprietary microcontroller
- Shifts control from hardware to embedded firmware
- Real time sensors for voltage, temperature, current/power
- Flexibility enables sophisticated algorithms, tuned for current operating conditions

Ref. R. Kumar, paper 3.2, ISSCC '09

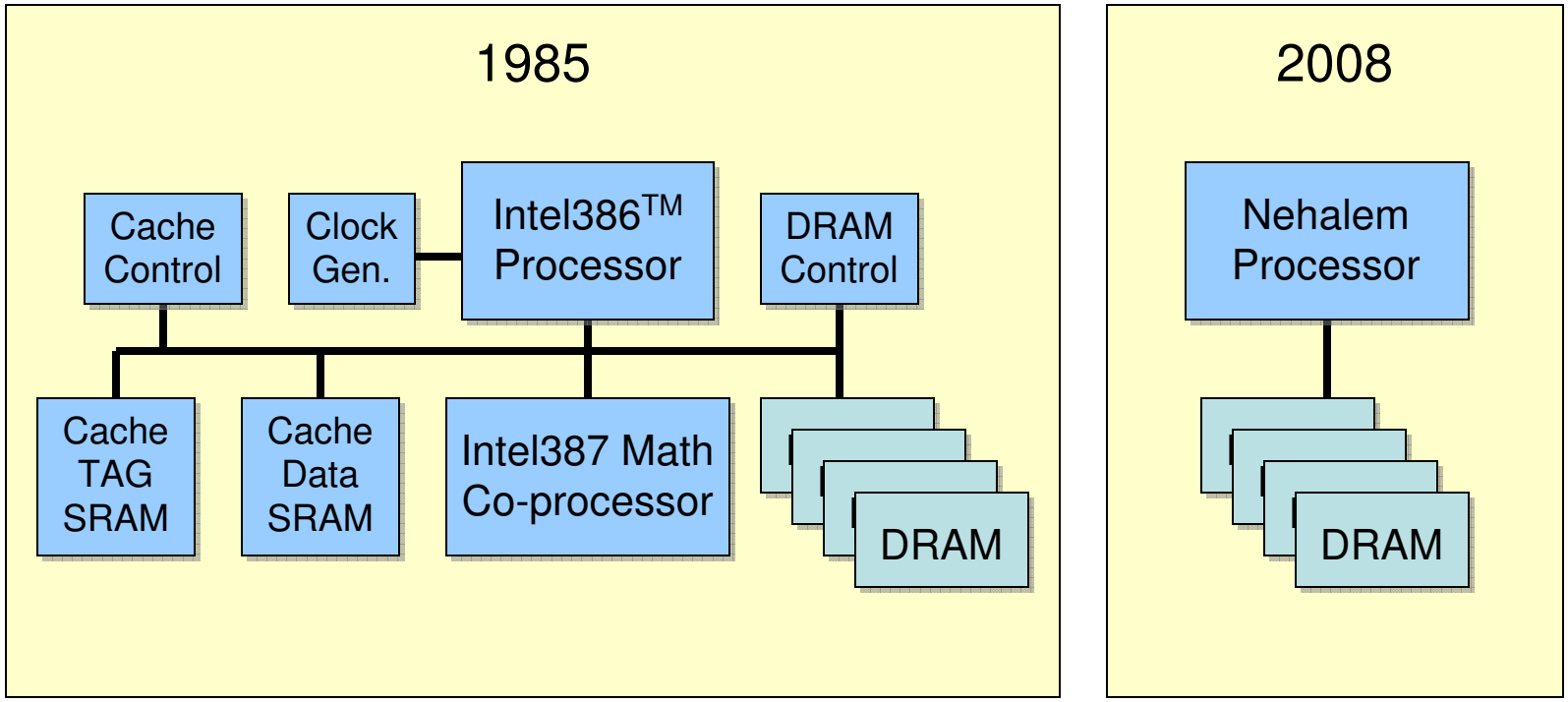
Adaptive Frequency System



- Adaptive PLL frequency
  - Higher frequency during voltage peaks
  - Lower frequency during voltage droops
- Up to 5% frequency improvement at same voltage
- Lower power at same frequency

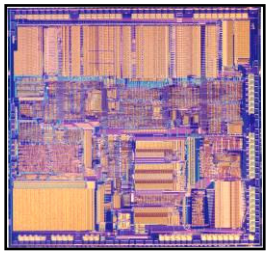
Ref. N. Kurd, VLSI Circuits '08

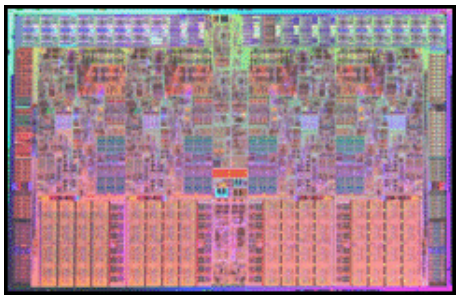
PC Platform Comparison



Modern microprocessors integrate many of the separate system components from past platforms

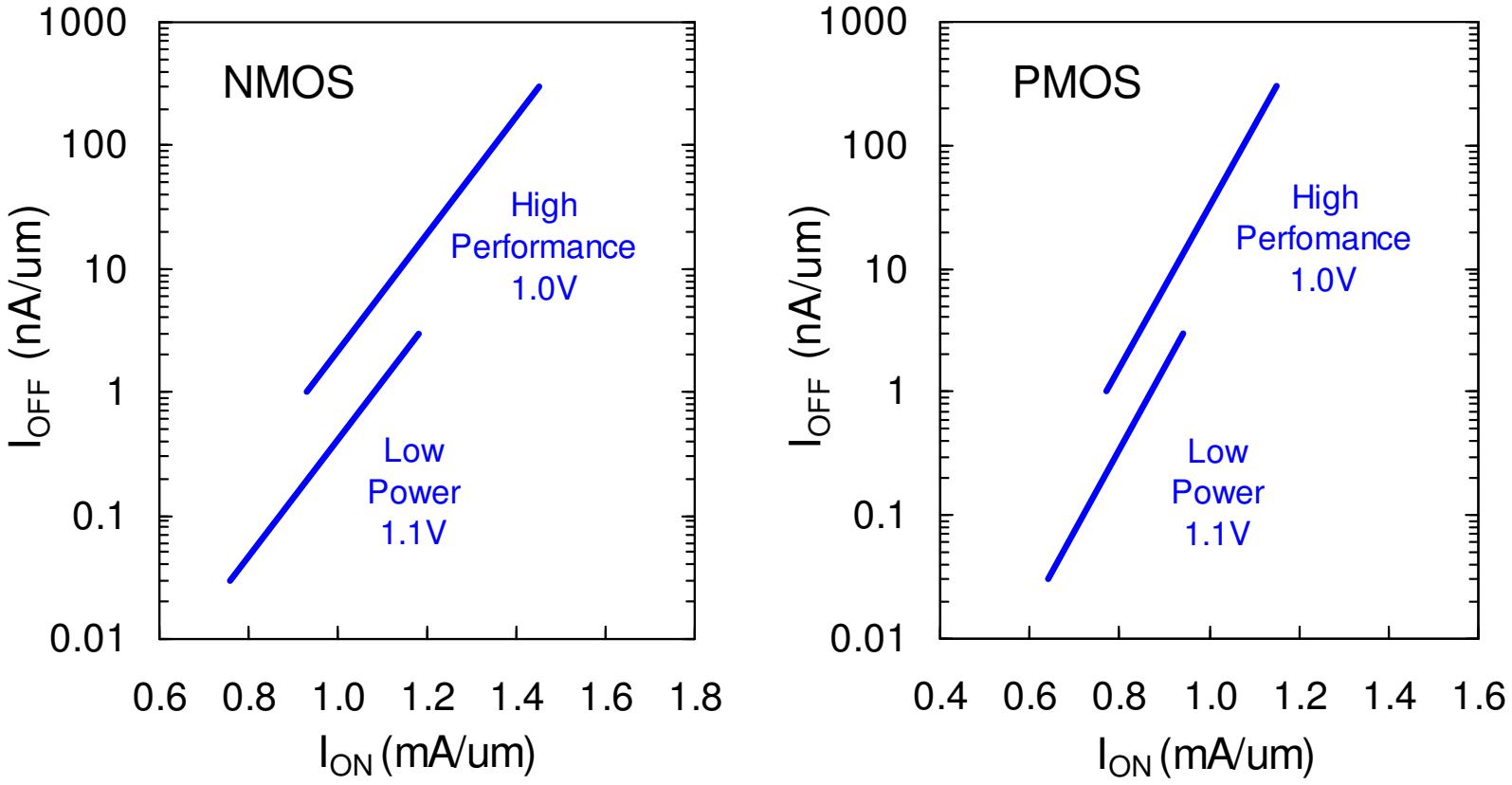
Microprocessor Evolution





	Intel386™	Nehalem
Transistor Count:	280 thousand	731 million
Frequency:	16 MHz	>3.6 GHz
# Cores:	1	4
Cache Size:	None	8 MB
I/O Peak Bandwidth:	64 MB/sec	50 GB/sec
Adaptive Circuits:	None	Sleep Mode Turbo Mode Power Gating Adaptive Frequency Clocking

45 nm SoC Transistors

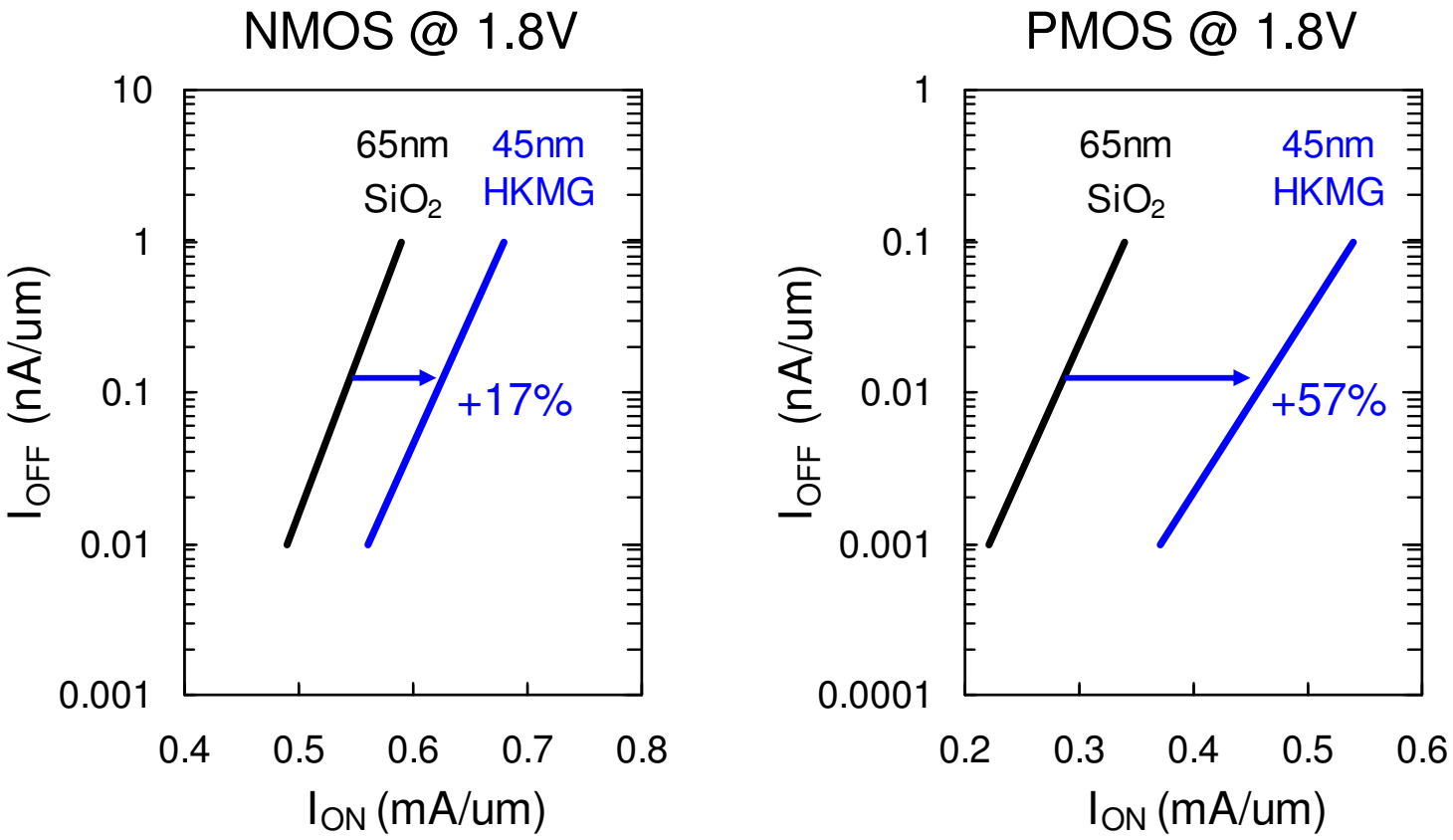


Wider range of transistor types provided for SoC:

High performance and low power

Ref. C. Jan, IEDM '08

45 nm SoC I/O Transistors



Wider range of transistor types provided for SoC:

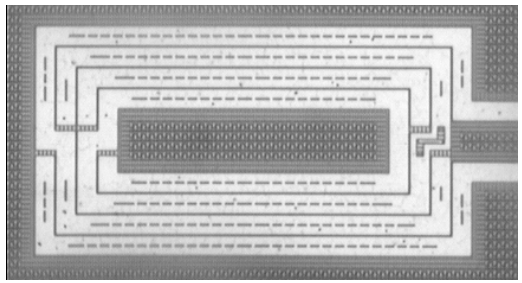
High speed, high voltage I/O

Ref. C. Jan, IEDM '08

Devices for SoC Analog Circuits

Passive Elements

- Precision resistor
- High Q varactor
- High Q inductor

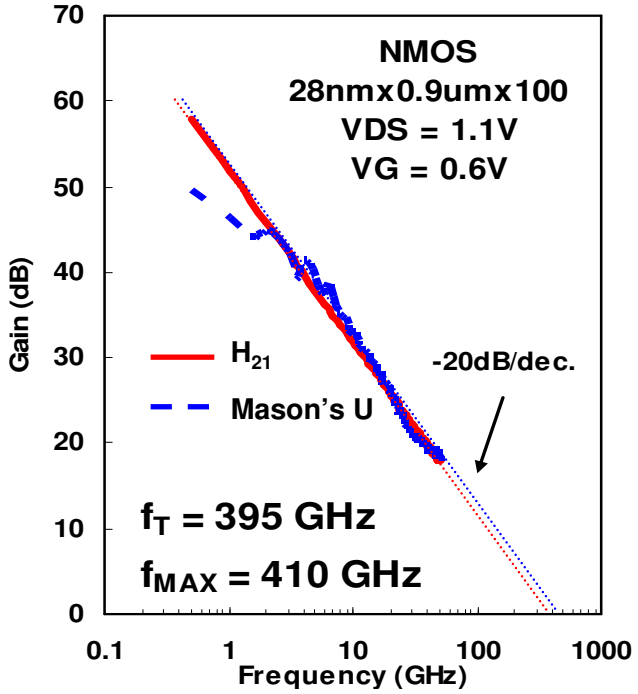


Active Elements

- RF CMOS

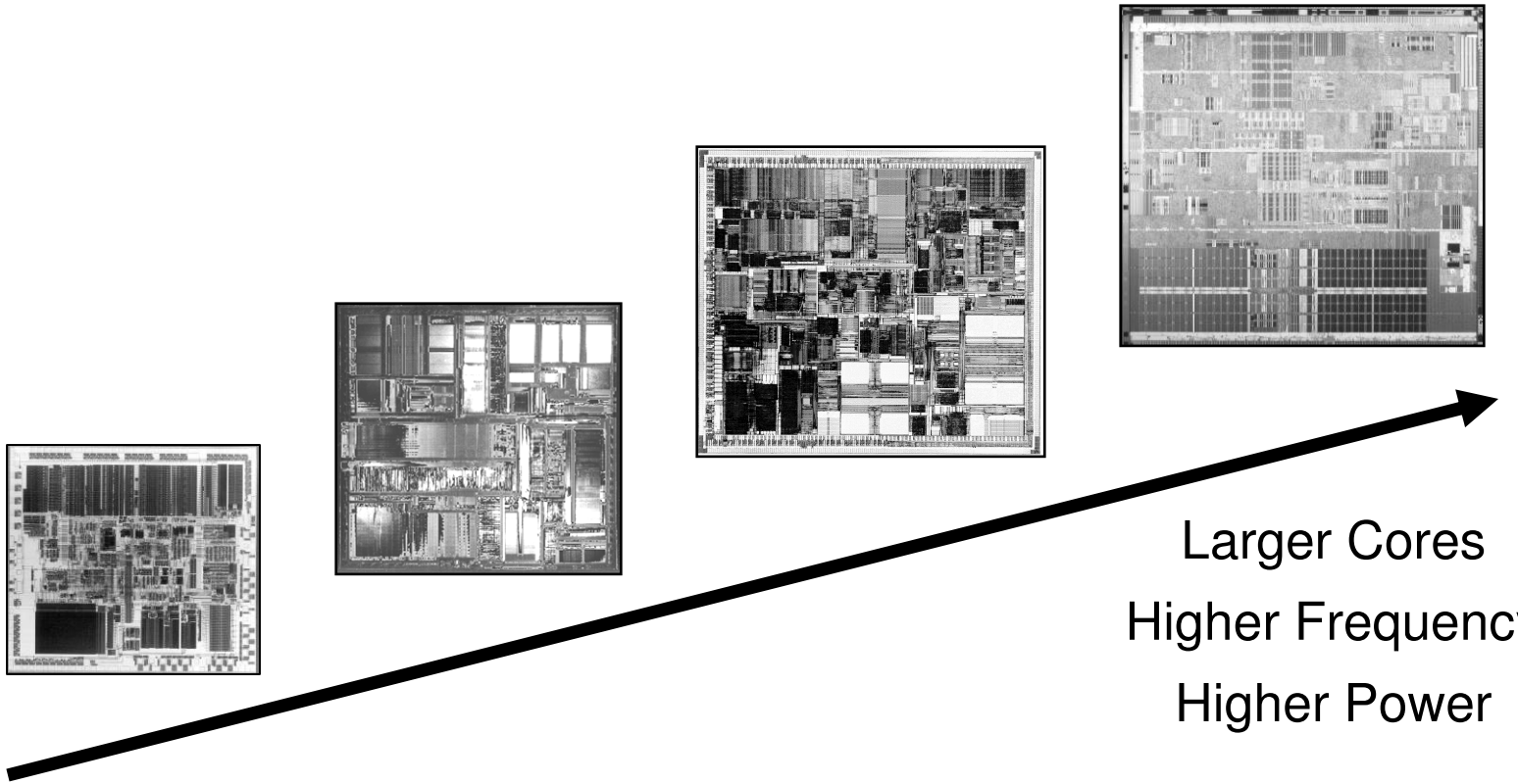
RF + Mixed Signal Circuits

- A to D, D to A converters
- RF transceiver
- LCPLL
- High speed I/O



Ref. C. Jan, IEDM '08

The Old Era of Microprocessor Scaling

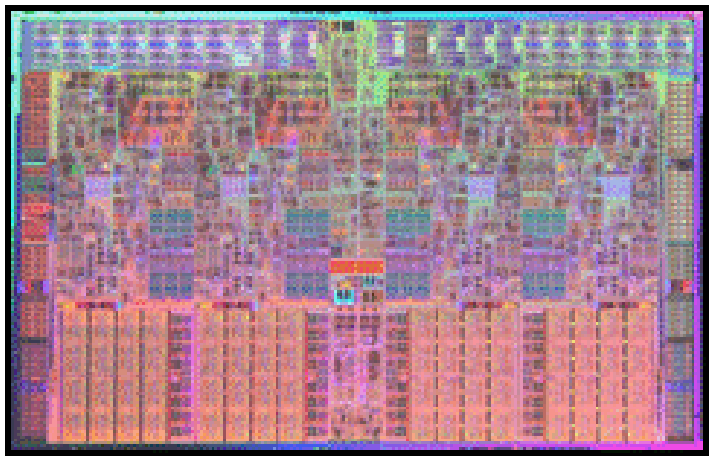


It has served us well for >30 years

The New Era of Microprocessor Scaling

Many-Core Multi-Core Multi-Function

System on a Chip



Avoiding the power wall requires a systemic approach from process technology through circuit design to micro-architecture to deliver products with power efficient performance

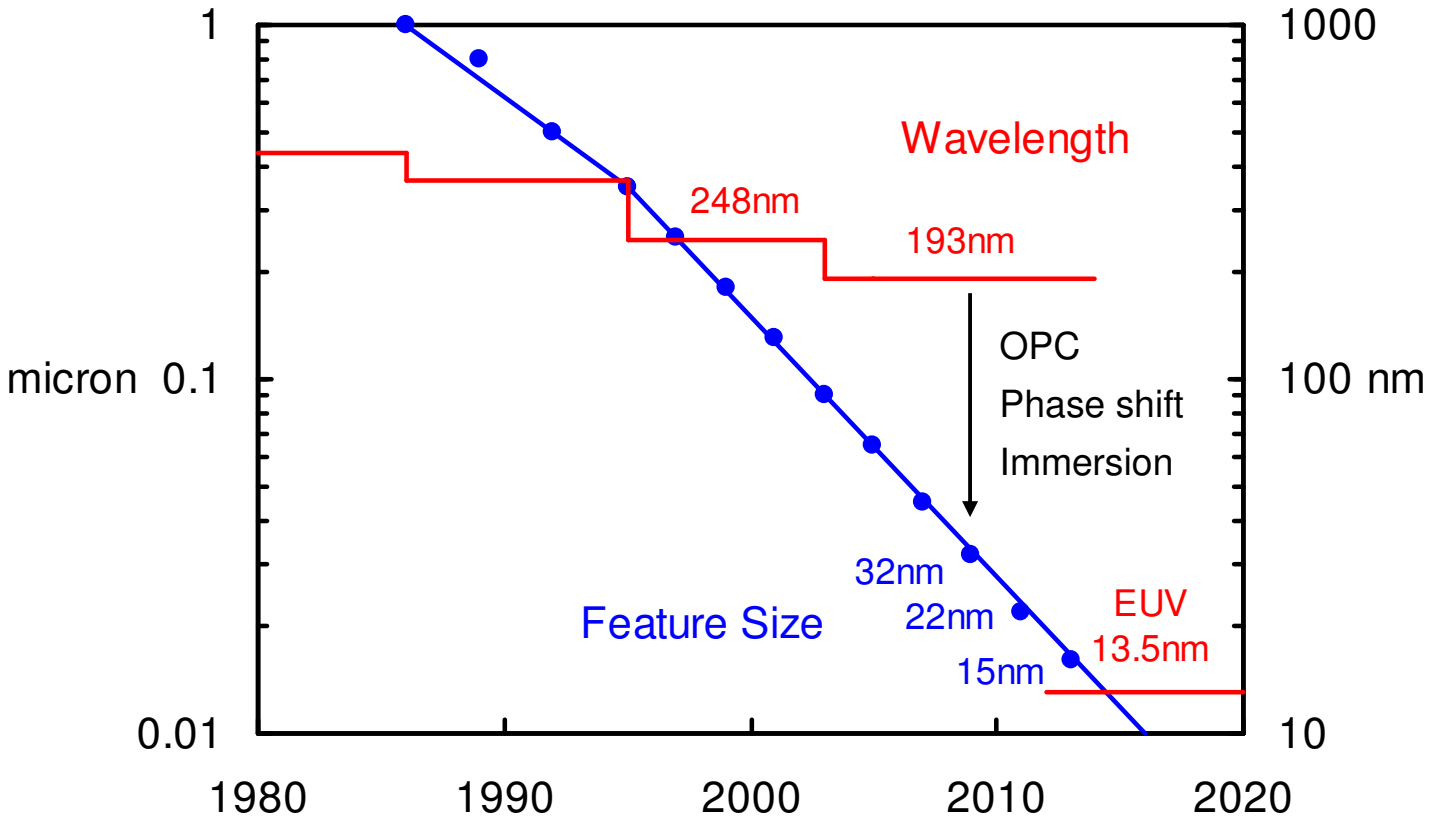
Outline

- Transistor Scaling
- Microprocessor Evolution
- Vision of the Future

Future Scaling Challenges

- Patterning ever-smaller features sizes
- Transistor and interconnect technologies that provide higher performance at lower power
- Continued voltage scaling for low power
- Integrating a wider range of device types for system-on-chip or system-in-package products

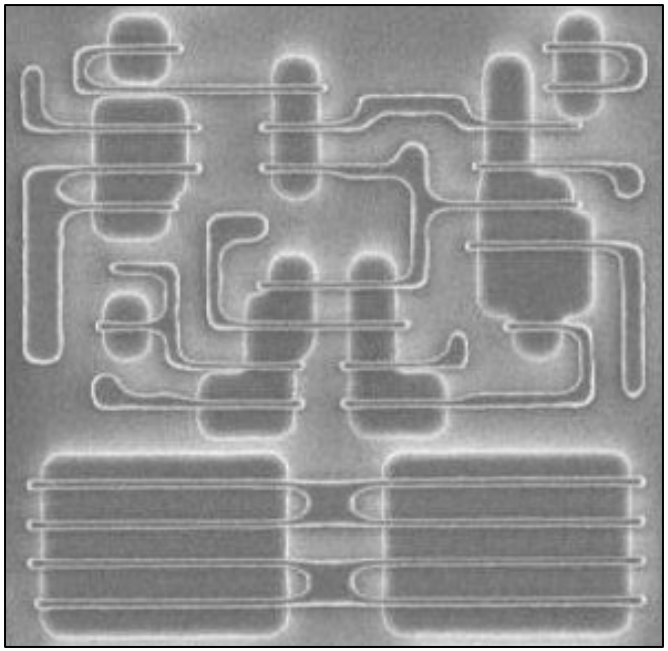
Lithography



193 nm enhancements got us to the 32 nm generation

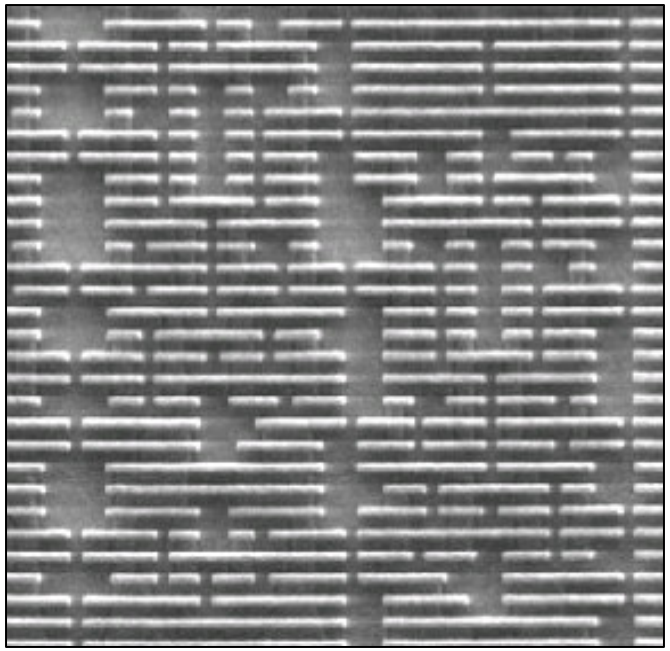
Layout Restrictions

65 nm Layout Style



- Bi-directional features
- Varied gate dimensions
- Varied pitches

32 nm Layout Style



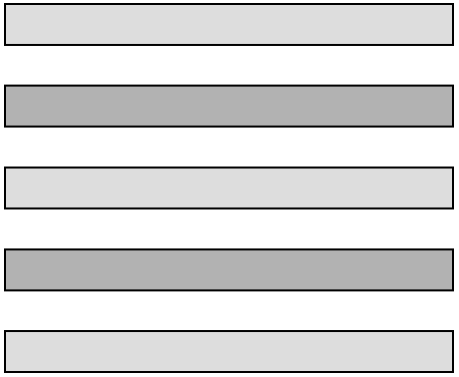
- Uni-directional features
- Uniform gate dimension
- Gridded layout

Lithography Options for Beyond 32 nm

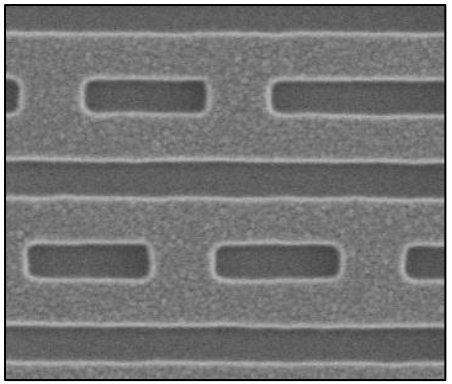
Double Patterning

- Pitch doubling
- Improved 2-D features

Pitch Doubling



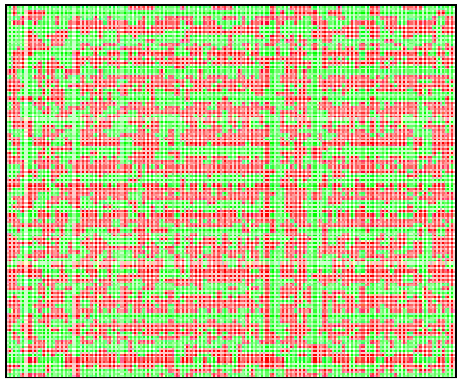
2-D Features



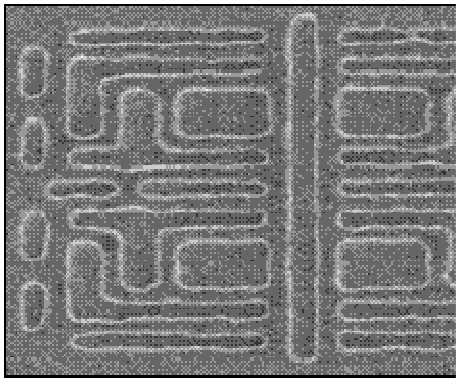
Computational Lithography

- Pixilated mask
- Existing 193 nm litho tools

Pixilated Mask

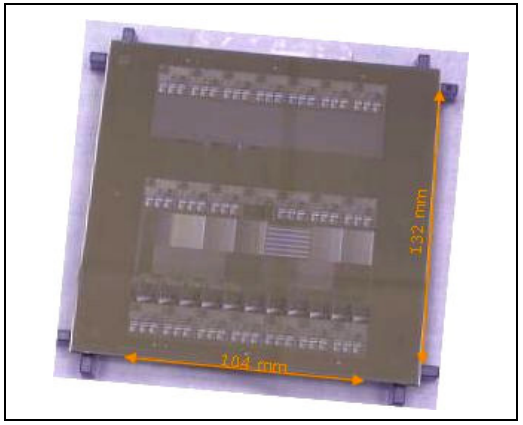


Printed Image

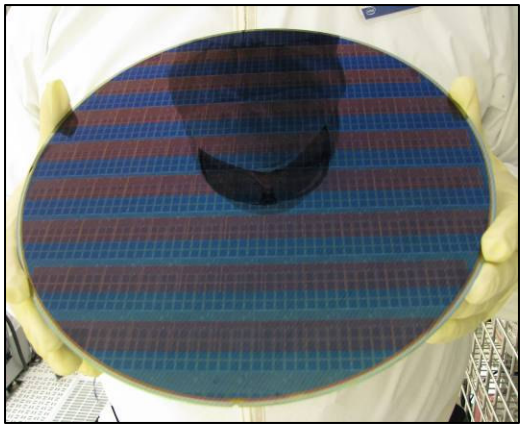


Extreme Ultraviolet Lithography

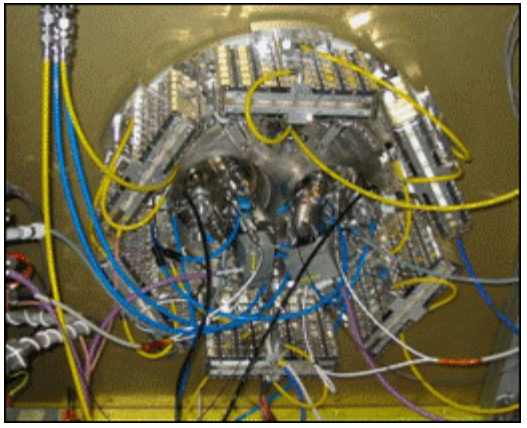
Cymer beta source



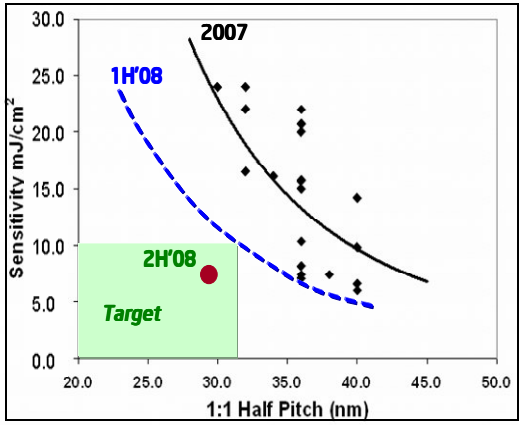
Intel EUV Mask



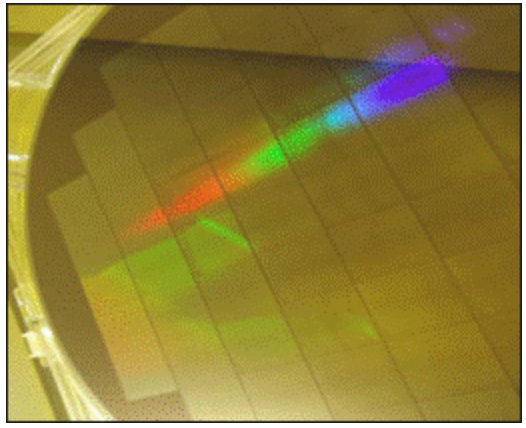
ASML ADT printed wafer



Philips beta source



Photoresist Development



Nikon EUV1 printed wafer

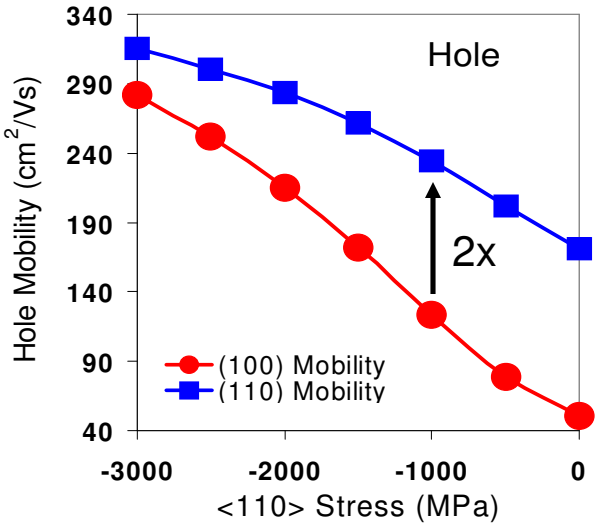
Continued progress towards EUV implementation

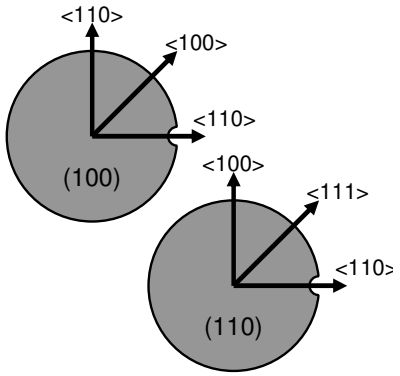
Transistor Options

Substrate Engineering

+ Increased p-channel mobility

? Impact on n-channel mobility





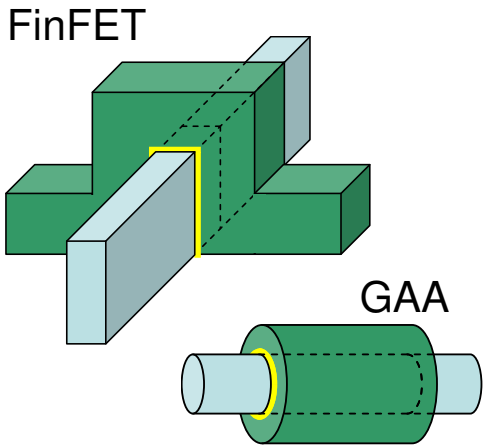
Multi-Gate FETs

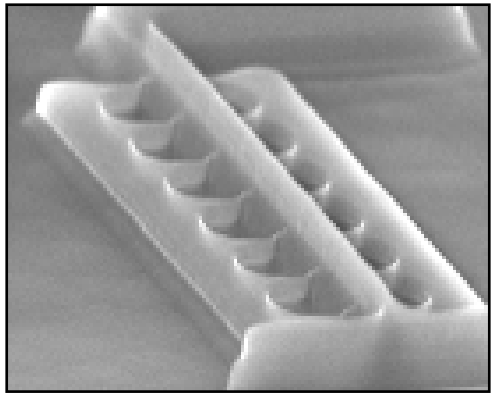
+ Improved electrostatics

+ Steeper sub-threshold slope

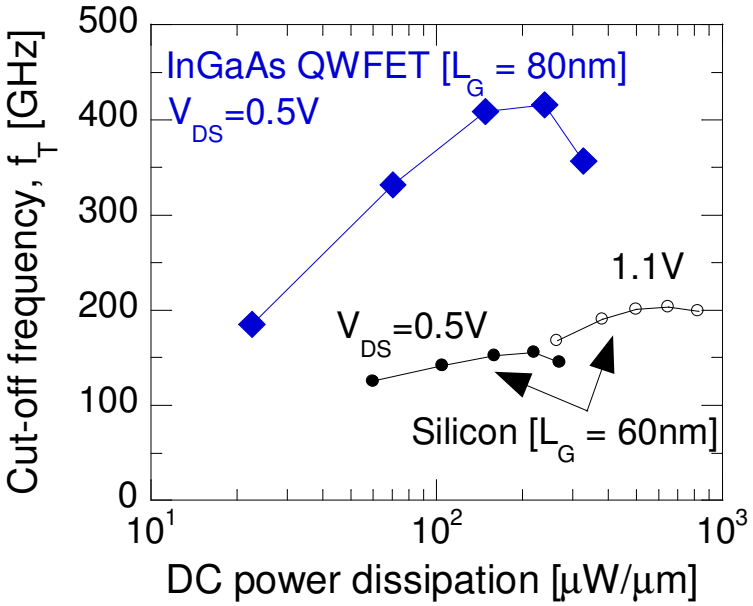
? Higher parasitic resistance

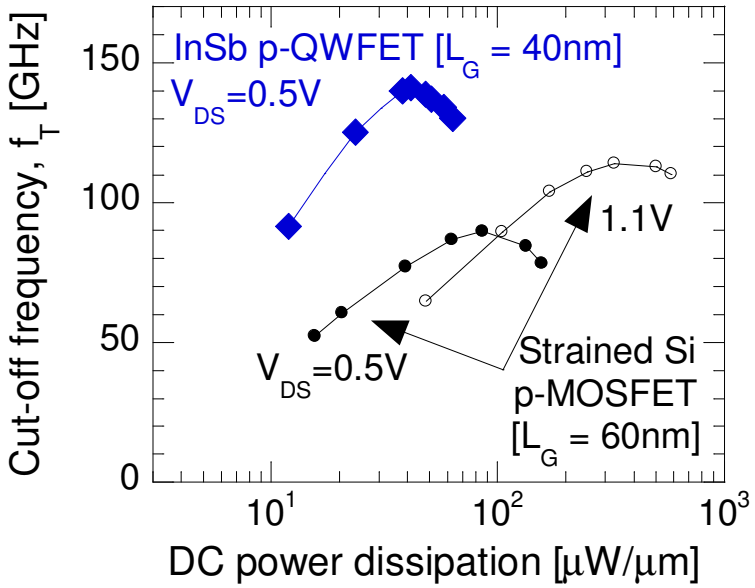
? Higher parasitic capacitance





III-V Transistor Options





InGaAs NMOS QWFET

Peak f_T > 400GHz at Vcc = 0.5V

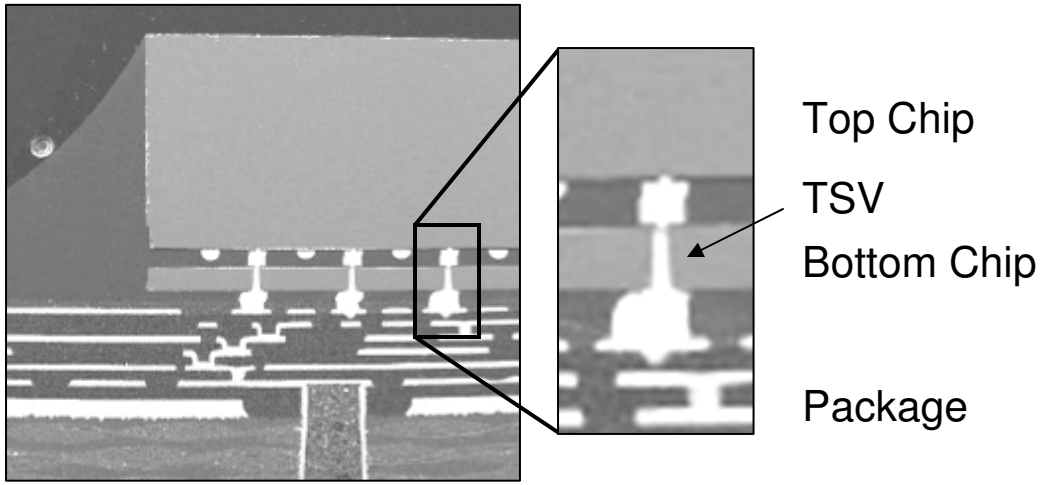
InSb PMOS QWFET

Peak f_T > 140 GHz at Vcc = -0.5V

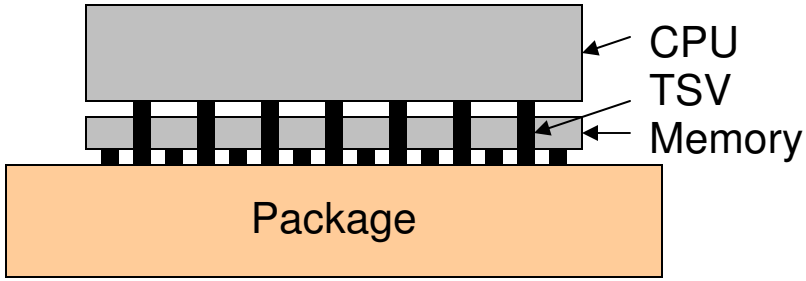
III-V materials for improved performance at low voltage

3-D Chip Stacking

+ High density chip-chip connections
+ Small form factor
+ Combine dissimilar technologies

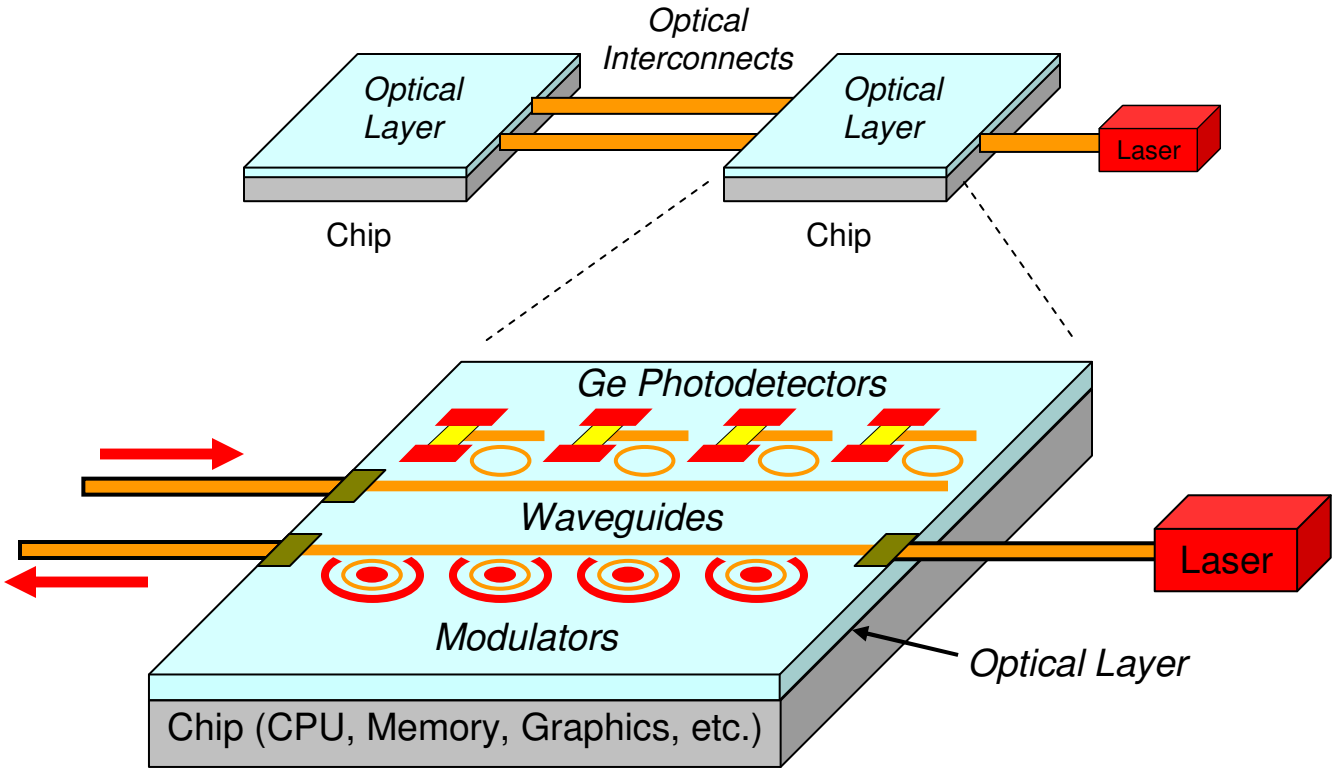


? Added cost
? Degraded power delivery, heat sinking
? Area impact on lower chip



3-D chip stacking using through-silicon vias

Optical Interconnects

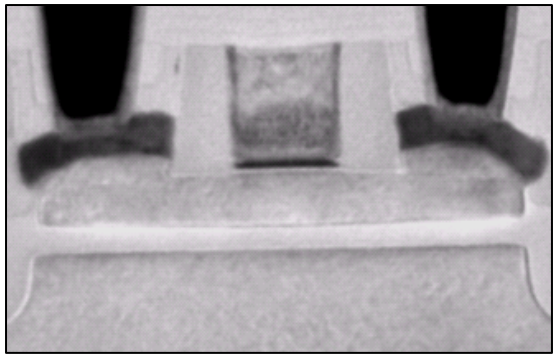


Nearer term: High bandwidth chip-chip interconnects

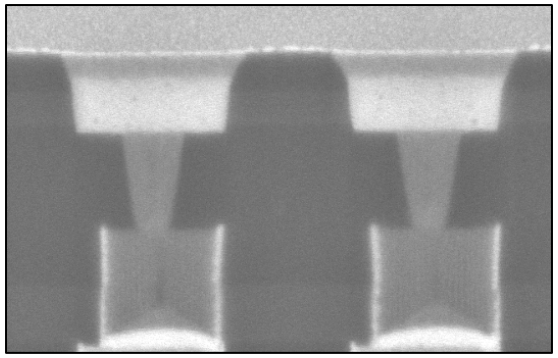
Longer term: On-chip interconnects

Ref. I. Young, paper 28.1, ISSCC '09

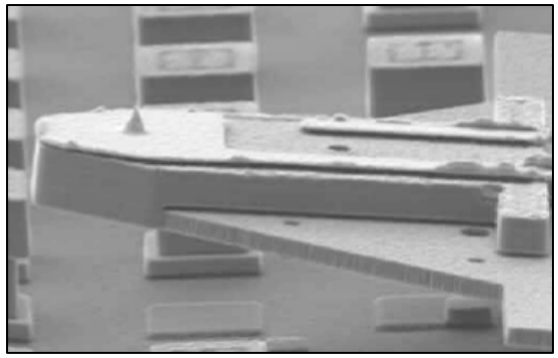
High Density Memory



Floating Body Cell



Phase Change Memory

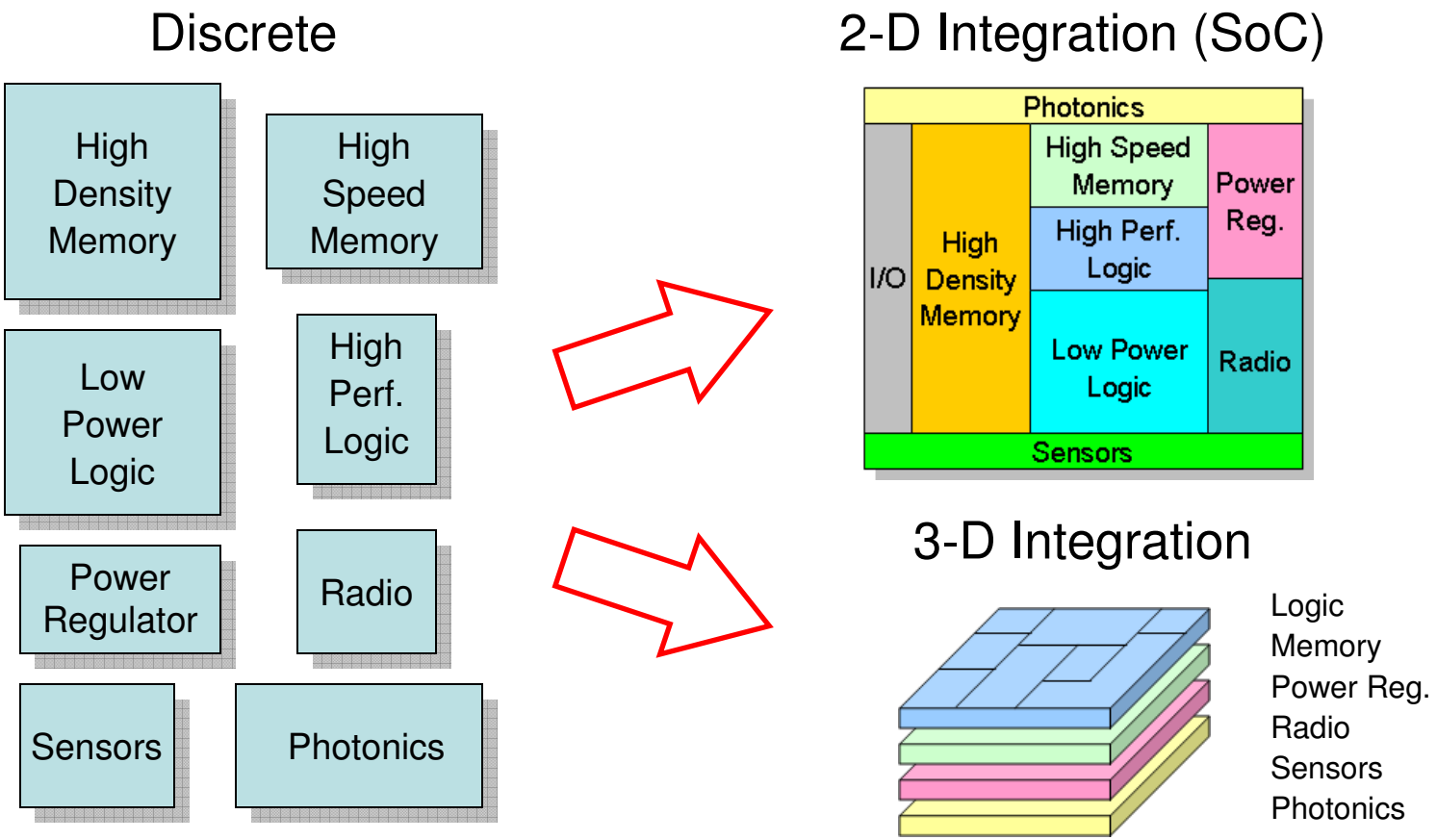


Seek and Scan Probe

Dense memory increasingly important

Several novel directions being explored

System Integration



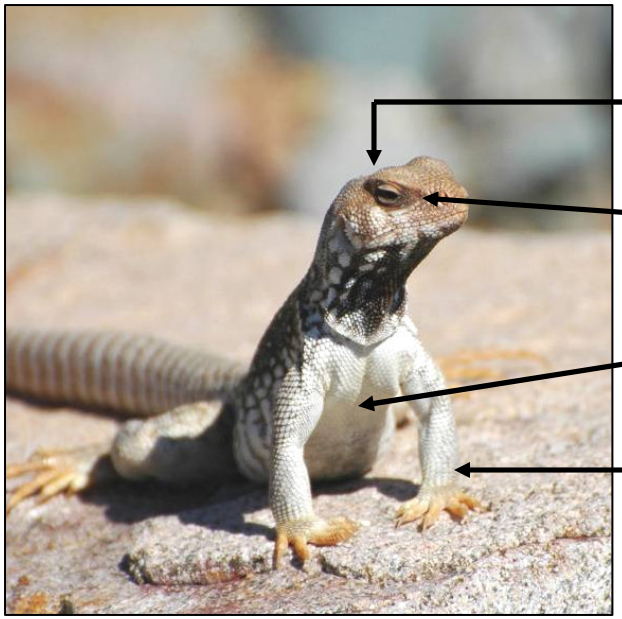
System integration needed for performance, power, form factor

Challenge is to integrate wider range of heterogeneous elements

Higher Level System Integration

Organic

Electronic

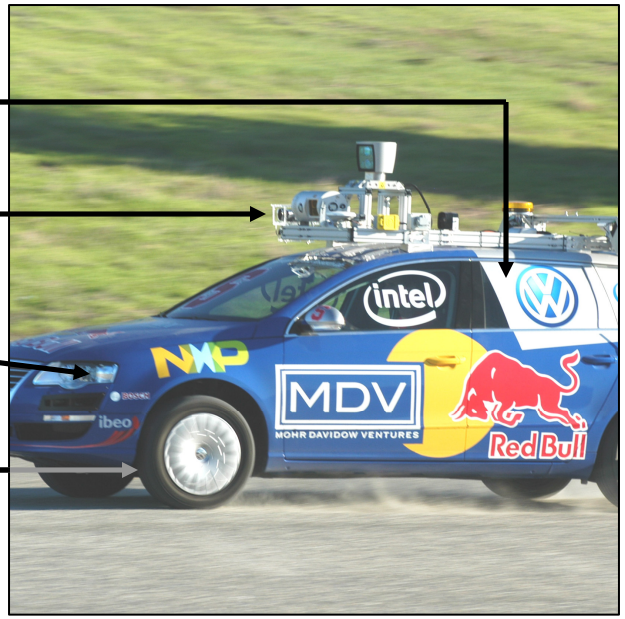


Computing

Sensors

Power Supply

Motion



Reptile

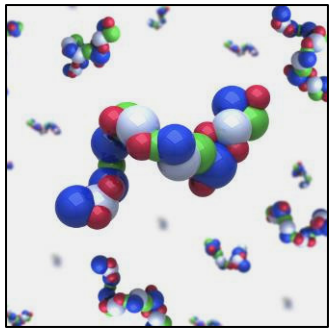
Autonomous Vehicle

Stanford entry
2007 DARPA challenge

We're trying to emulate nature's capabilities

Evolutionary Comparison

Organic

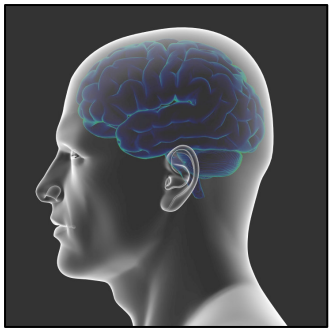


Complex Molecule

Single-Cell Organism

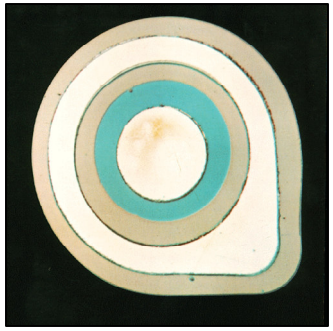
Multi-Cell Organism

Reptile

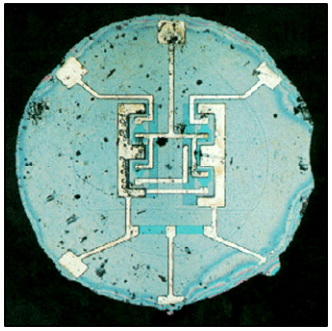


Human

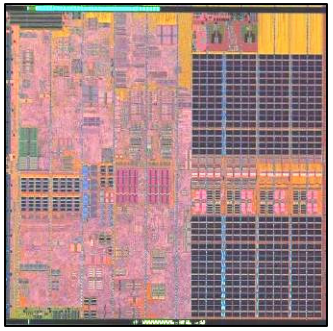
Electronic



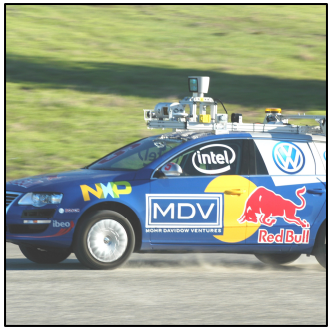
Transistor



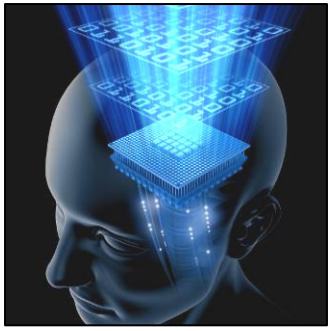
Integrated Circuit



Microprocessor PC



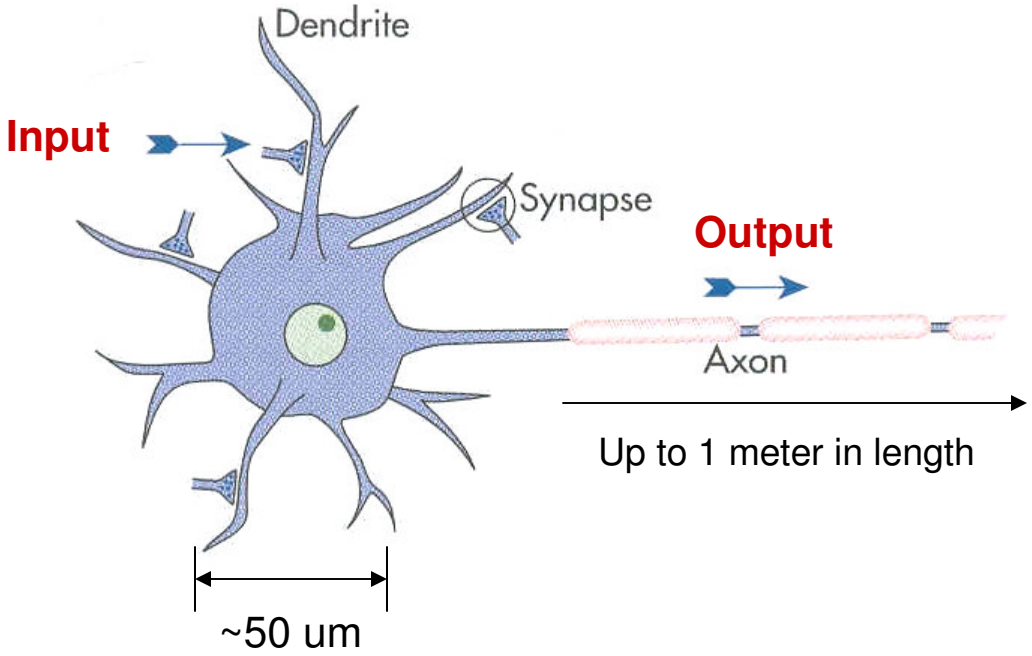
Autonomous Vehicle



Robot

What did nature have to "invent" to evolve to higher forms?

Brain Neuron

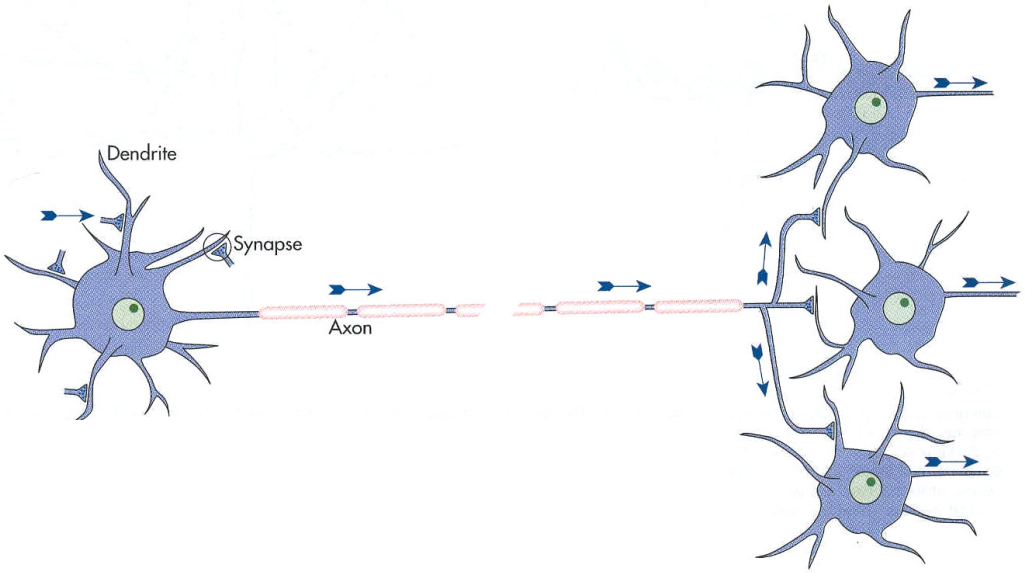


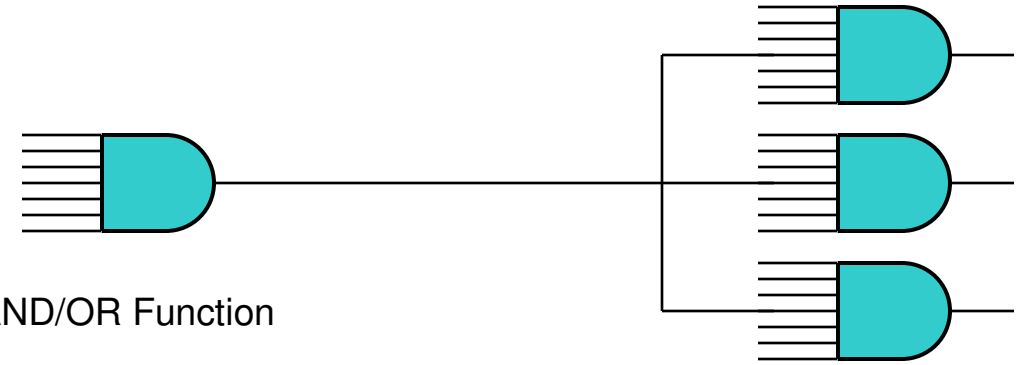
	Neuron	Transistor
Charge carrier:	Ions	Electrons
Voltage swing:	100 mV	1.0 V
Threshold voltage:	10-20 mV	~300 mV

Nature is a master of low power operation

Neuron image from J. Nolte [36]

Organic vs. Electronic Circuits





AND/OR Function

FI, FO ~1000

Operates ~100 Hz

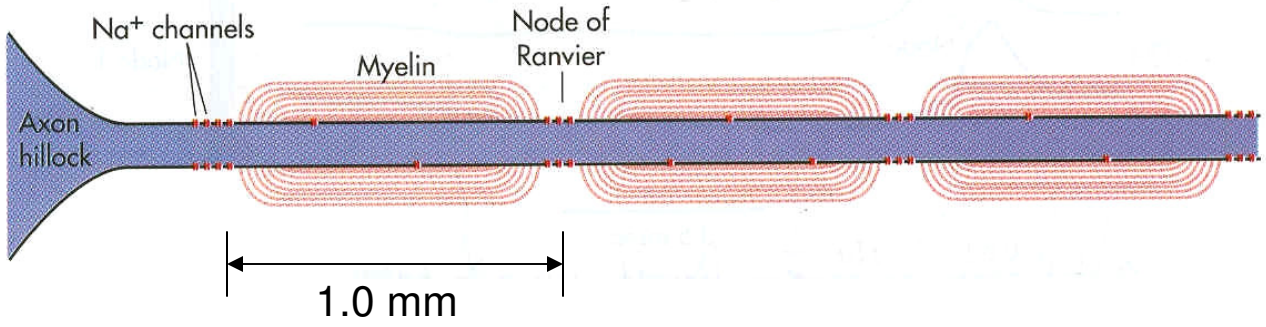
Brain circuits are slow but massively parallel

Neuron image from J. Nolte [36]

Organic vs. Electronic Interconnects

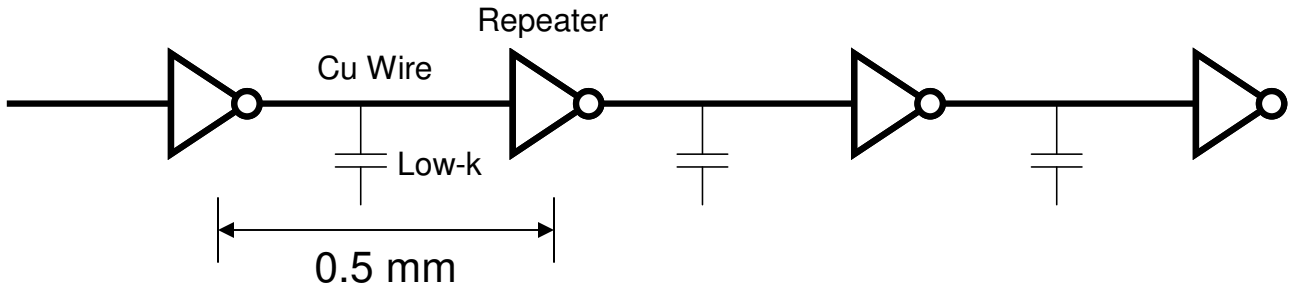
Myelinated Axon

25 m/sec →



Cu + Low-k + Repeaters

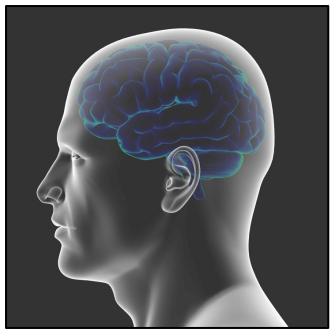
>10^7 m/sec →



Myelin coating improves axon signal speed ~10x, but still slow

Axon image from J. Nolte [36]

Organic vs. Electronic Systems



# Devices:	10^{11} Neurons 10^{14} Synapses ✓	$>10^{8}$ CPU Transistors 10^{11} System Total
Input Devices:	Eyes, Ears, Taste, Touch, Smell ✓	Keyboard, Radio, USB Port
Operating Freq:	100 Hz	>2 GHz ✓
Power:	20 Watts ✓	40 Watts

We have a way to go and much to learn

Conclusion

- Moore's Law continues, but the formula for success is changing
  - New materials and device structures are needed to continue scaling
  - Circuit design and micro-architecture innovations focus more on power efficiency
- System level integration is increasingly important
  - Success will be determined by ability to integrate a wider and more heterogeneous set of components
- Organic evolution has given us some clues for effective higher level system integration
  - Low power operation
  - Massive parallelism
  - Integrated sensors